공학박사 학위논문

기계학습을 활용한
지도 일반화 개선 방안 연구

-건물과 도로객체의 선택적 삭제를 중심으로-

A Study on Improvement of Map Generalization Using Machine Learning -Focusing on Selective Omission of Building and Road Data-

2020년 7월

서울대학교 대학원

건설환경공학부

이 재 은

국문초록

현재 우리나라의 1:25,000 수치지형도는 1:5,000 수치지형도를 축소 편집하여 생성하고 있다. 축소 편집은 대축척 지도로부터 소축척 지도를 만드는 과정이며, 이 과정에서 필연적으로 지도 일반화(map generalization) 기법을 적용하게 된다. 그동안의 지도 일반화는 객체의 기하학적 특징들을 활용하여 일반화하는 기하학적 일반화 방법, 혹은 규칙기반(rule-based) 방법이 주류를 이루어 왔다. 현재 우리나라의 축소 편집 과정은 축소 편집 관련 규정에 따르는 일종의 규칙기반 방법을 통해 수행되고 있다고 볼 수 있다. 하지만, 규정집의 내용이 구체적이지 않은 부분들이 다수 존재하여 편집자의 주관이 개입될 여지가 많다. 축소 편집 시에 제작자의 주관이 개입될수록 소축척 지도 품질의 일관성을 담보해 주지 못하며, 편집자 개인의 역량에 따라 일반화의 품질이 좌우된다는 단점이 있다.

지도 일반화에 관한 연구에서도 사람의 개입이 지도 일반화의 결과물을 일관적이지 못하게 한다는 문제가 꾸준히 제기되고 있다. 이에 따라 지도 일반화의 연구 흐름은 자연스럽게 사람의 개입을 최소화하고 자료 취득 및 처리 공정을 자동화함으로써 지도 일반화 품질의 일관성을 담보할 수 있는 방향으로 진행되어왔다. 그러나 이에 앞서 사람의 개입이 지도 일반화 품질에 어떠한 영향을 얼마나 주고 있는지에 대해 정량적으로 밝혀진 연구사례나 실증사례는 매우 부족한 현실이며, 구축된 지도 일반화 결과물의 분석을 통해 기존의 규정을 보완하는 등의 활용 방안을 제시하려는 시도 또한 미흡한 편이다.

본 연구에서는 사람의 개입으로 인해 발생하는 일반화 품질의 차이를

기계학습 방법을 적용하여 정량화하고, 나아가 이를 활용하여 일반화된 지도의 품질 향상 방안을 제시하고자 한다. 이를 위해 1:5,000 수치지형도에서 1:25,000 수치지형도로의 축소 편집 시 건물과 도로의 선택적 삭제 여부를 예측할 수 있는 기계학습 모델을 생성하고 학습된 모델을 서로 다른 여섯 명의 지도 제작자가 제작한 지역에 적용하여 건물과 도로 객체의 선택적 삭제 여부에 대한 예측률을 측정하였다. 측정된 예측률 간의 차이와 그 양상을 분석함으로써 지도 제작자 간의 편집 방법에 있어서 유의미한 차이가 있음을 밝히고자 하였다.

이를 위해 학습 모델의 성능평가를 위해 학습 모델에 사용된 네 개의 알고리듬 - 의사결정 나무(decision tree), k-최근접 이웃(k-nearest neighbor), SVM(Support Vector Machine), 인공신경망(Artificial Neural Network, ANN) - 별로 건물과 도로에 대해 각각 예측률을 측정하였다. 또한, 각각 생성된 모델을 6개 실험지역에 적용하여 예측률을 측정하였고 크루스칼 왈리스 검정을 통해 예측률 간의 차이가 통계적으로 유의미한 수준임을 볼 수 있었다.

이 과정에서 대상 지역의 특징에 따라 정확도의 차이가 발생할 수 있으므로 같은 제작자가 편집한 서로 다른 지역에 대한 정확도를 측정하고 통계 검증을 통해 지역의 특징이 제작자 간 차이에 얼마나 영향을 미치는지에 대하여 분석하였다. 그 결과 건물의 경우 지역별로 드러난 정확도의 차이가 통계적으로는 유의미한 수준이 아니었으나, 도로의 경우 일부 지역에서 유의미한 차이가 나타났다. 그러나 지역별로 드러난 차이보다 제작자별로 나타난 차이가 더 크게 나타났으며, 이는 도로객체에 대해서도 제작자별 차이가 발생할 수 있다고 해석할 수 있다. 정성적(시각적) 분석 결과, 건물의 경우 도시지역의 소건물들에서, 도로의 경우 진입로 등의 소로에서 객체의 선택적 삭제에 제작자별 차이가 드러나는 것을

발견할 수 있었다.

또한, 기계학습 기법의 지도 일반화 분야에서의 활용 방안 모색을 위해 도시와 비 도시 지역에 대해 각각 기계학습 모델을 생성하고 예측률을 측정하였다. 그 결과 기계학습 기법을 통해 객체의 선택적 삭제 과정에서 주된 영향을 주는 속성들과 객체 선택에 필요한 설정값들을 확인할 수 있었다. 이를 통해 기계학습 알고리듬이 지도 일반화 과정에서 축소 편집 규정을 보완할 수 있으며, 특징별로 기계학습 기법을 통해 객체의 선택적 삭제에 기초적 활용이 가능함을 보였다.

기계학습 기법의 적용을 통해 지도 제작자 간의 편차를 정량화하는 것에서 나아가 지도 제작 또는 지도 일반화 과정에 대한 검수 프로그램으로써 활용도 기대할 수 있다. 또한, 대상 객체를 확대하고 알고리듬의 예측률을 상승시키면 현재의 지도 축소 편집 과정을 제작자의 개입 없이 자동화가 가능할 것으로 기대된다.

주요어 : 지도 일반화, 지도 갱신, 다축척 데이터베이스, 기계학습, 수치지형도

학 번 : 2011-30271

목 차

1. 서론 ····· 1

1.1. 연구 배경 및 목적 ····· 1

1.2. 연구 동향 ····· 10

1.2.1. 건물과 도로의 일반화 기법 ····· 10

1.2.2. 기계학습 기법을 적용한 연구 ····· 16

1.2.3. 현재의 지도 제작 및 일반화 과정 ····· 20

1.2.4. 소결 ····· 29

1.3. 연구의 범위 및 방법 ····· 30

2. 기계학습 알고리듬 ····· 35

2.1. 기계학습 개요 ····· 35

2.2. 사용된 기계학습 알고리듬 ····· 40

2.2.1. 의사결정 나무 ····· 41

2.2.2. k-최근접 이웃 ····· 43

2.2.3. SVM ····· 46

2.2.4. 인공신경망 ····· 49

2.3. 기계학습 알고리듬을 위한 데이터 생성 ····· 53

2.3.1. 실험 데이터 및 전체 실험 순서 ····· 53

2.3.2. 훈련 데이터 생성 ····· 55

3. 실험 및 결과 ····· 64

3.1. 건물과 도로에 기계학습 적용 ····· 64
3.1.1. 건물에 적용 ····· 64
3.1.2. 도로에 적용 ····· 70
3.2. 학습 모델의 생성 및 성능 평가 ····· 73
3.3. 모델 적용을 통한 제작자 간 차이의 정량화 ····· 80
3.3.1. 건물에 대한 성능평가 ····· 81
3.3.2. 도로에 대한 성능평가 ····· 87
3.3.3. 지역 간의 차이에 대한 평가 ····· 92
3.4. 도시와 비 도시 지역에 적용 ····· 99
3.5. 소결 ····· 106

4. 결론 및 고찰 ····· 108

참고문헌 ····· 112
Abstract ····· 128

표 목 차

[표 1-1] 지도 일반화를 위한 필수 연산자들 ····· 3
[표 1-2] 수치지형도 1.0과 수치지형도 2.0 ····· 22
[표 1-3] 축소 편집 시 도로의 표시에 관한 규정 ····· 26
[표 1-4] 축소 편집 시 건물의 표시에 관한 규정 ····· 27
[표 2-1] 사용된 기계학습 알고리듬들의 특징 ····· 40
[표 2-2] 제작자별 건물과 도로의 객체 수 ····· 55
[표 2-3] 사용된 입력 속성(건물) ····· 57
[표 2-4] 사용된 입력 속성(도로) ····· 63
[표 3-1] 건물의 실험 데이터 구조 예시 ····· 65
[표 3-2] 의사결정 나무 알고리듬의 오차 행렬(모델-건물) ····· 73
[표 3-3] k-최근접 이웃 알고리듬의 오차 행렬(모델-건물) ····· 73
[표 3-4] SVM 알고리듬의 오차 행렬(모델-건물) ····· 74
[표 3-5] 인공신경망 알고리듬의 오차 행렬(모델-건물) ·· 74
[표 3-6] 의사결정 나무 알고리듬의 오차 행렬(모델-도로) ····· 76
[표 3-7] k-최근접 이웃 알고리듬의 오차 행렬(모델-도로) ····· 77
[표 3-8] SVM 알고리듬의 오차 행렬(모델-도로) ·····

…… 77
[표 3-9] 인공신경망 알고리듬의 오차 행렬(모델-도로) ·· 77
[표 3-10] 건물의 모델 및 각 지도 제작자별 예측률 …… 81
[표 3-11] 건물의 모델과 실험 대상 지역의 크루스칼 왈리스 검정 결과 …… 82
[표 3-12] 건물 면적별 예측률 - 제작자 A …… 83
[표 3-13] 건물 면적별 예측률 - 제작자 B …… 83
[표 3-14] 건물 면적별 예측률 - 제작자 C …… 83
[표 3-15] 건물 면적별 예측률 - 제작자 D …… 83
[표 3-16] 건물 면적별 예측률 - 제작자 E …… 84
[표 3-17] 건물 면적별 예측률 - 제작자 F …… 84
[표 3-18] 도로의 모델 및 각 지도 제작자별 예측률 …… 88
[표 3-19] 도로의 모델과 실험 대상 지역의 크루스칼 왈리스 검정 결과 …… 88
[표 3-20] 도로 폭에 따른 예측률 - 제작자 A …… 89
[표 3-21] 도로 폭에 따른 예측률 - 제작자 B …… 89
[표 3-22] 도로 폭에 따른 예측률 - 제작자 C …… 89
[표 3-23] 도로 폭에 따른 예측률 - 제작자 D …… 89
[표 3-24] 도로 폭에 따른 예측률 - 제작자 E …… 90
[표 3-25] 도로 폭에 따른 예측률 - 제작자 F …… 90
[표 3-26] 건물에서의 지역 간의 차이 - 제작자 A …… 93
[표 3-27] 제작자 A의 건물에 대한 크루스칼 왈리스 검정 결과 …… 94

[표 3-28] 건물에서의 지역 간의 차이 - 제작자 B ····· 94
[표 3-29] 제작자 B의 건물에 대한 크루스칼 왈리스 검정 결과 ····· 94
[표 3-30] 건물에서의 지역 간의 차이 - 제작자 C ····· 95
[표 3-31] 제작자 C의 건물에 대한 크루스칼 왈리스 검정 결과 ····· 95
[표 3-32] 도로에서의 지역 간의 차이 - 제작자 A ····· 96
[표 3-33] 제작자 A의 도로에 대한 크루스칼 왈리스 검정 결과 ····· 96
[표 3-34] 도로에서의 지역 간의 차이 - 제작자 B ····· 97
[표 3-35] 제작자 B의 도로에 대한 크루스칼 왈리스 검정 결과 ····· 97
[표 3-36] 건물에서의 지역 간의 차이 - 제작자 C ····· 97
[표 3-37] 제작자 C의 도로에 대한 크루스칼 왈리스 검정 결과 ····· 98
[표 3-38] 도시와 비 도시 학습모델에서의 의사결정나무 입력변수 ····· 100
[표 3-39] 도시와 비 도시 학습모델에서의 k-최근접 이웃 입력변수 ····· 101
[표 3-40] 도시와 비 도시에서의 예측률 ····· 101
[표 3-41] 분석 결과에 따른 규정 개정안 ····· 104

그 림 목 차

[그림 1-1] 모델 일반화와 지도학적 일반화 ····· 2
[그림 1-2] ArcGIS의 ModelBuilder ····· 14
[그림 1-3] DeepOSM의 결과로 추출된 도로 ····· 19
[그림 1-4] 수치지형도 수시수정 주요 내용 ····· 21
[그림 1-5] 국가기본도 수정 사업 위치도 ····· 31
[그림 1-6] 논문 구성 및 흐름도 ····· 33
[그림 2-1] 기계학습 알고리듬의 종류 ····· 35
[그림 2-2] MNIST 데이터 세트 ····· 36
[그림 2-3] 비지도 학습의 예시 - 군집화 ····· 38
[그림 2-4] 강화학습 알고리듬 ····· 39
[그림 2-5] k-최근접 이웃 알고리듬의 예시 ····· 44
[그림 2-6] SVM의 최대 마진 초평면과 서포트 벡터 ····· 46
[그림 2-7] 뉴런의 값 계산 과정 ····· 50
[그림 2-8] 실험 순서도 ····· 53
[그림 2-9] 실험 데이터 생성 과정 ····· 54
[그림 2-10] 기계학습을 위한 입/출력 데이터 생성 ····· 56
[그림 2-11] 1:5,000 건물 레이어(좌)와 1:25,000 건물 레이어(우) ····· 58
[그림 2-12] 건물의 클래스 부여 과정 ····· 59
[그림 2-13] 1:5,000 도로중심선 레이어(좌)와 1:25,000

도로중심선 레이어(우) ····· 60
[그림 2-14] 수치지형도 도로중심선 재구조화 과정 ····· 61
[그림 2-15] 버퍼 폴리곤을 활용한 도로중심선 레이어 분류 과정 ····· 62
[그림 3-1] 건물에 대해 생성된 의사결정 나무 ····· 67
[그림 3-2] k 값에 따른 오차율 변화(건물) ····· 69
[그림 3-3] 도로에 대해 생성된 의사결정 나무 ····· 71
[그림 3-4] k 값에 따른 오차율 변화(도로) ····· 72
[그림 3-5] 건물을 대상으로 한 알고리듬별 예측률 ····· 75
[그림 3-6] 도로를 대상으로 한 알고리듬별 예측률 ····· 78
[그림 3-7] 제작자 간 차이의 정량화 과정 ····· 80
[그림 3-8] 제작자 A의 건물에 관한 결과 - 음영: 예측 결과의 오답 ····· 85
[그림 3-9] 제작자 B의 건물에 관한 결과 - 음영: 예측 결과의 오답 ····· 86
[그림 3-10] 제작자 B의 10000m^2 이상 건물에서의 오차 - 음영: 삭제됨 ····· 87
[그림 3-11] 제작자 B의 단지 내 도로에 대한 오차 결과 91
[그림 3-12] 제작자 E의 단지 내 도로에 대한 오차 결과 91
[그림 3-13] 도시와 비 도시 지역의 기계학습 모델에 대한 분석 순서도 ····· 100

1. 서 론

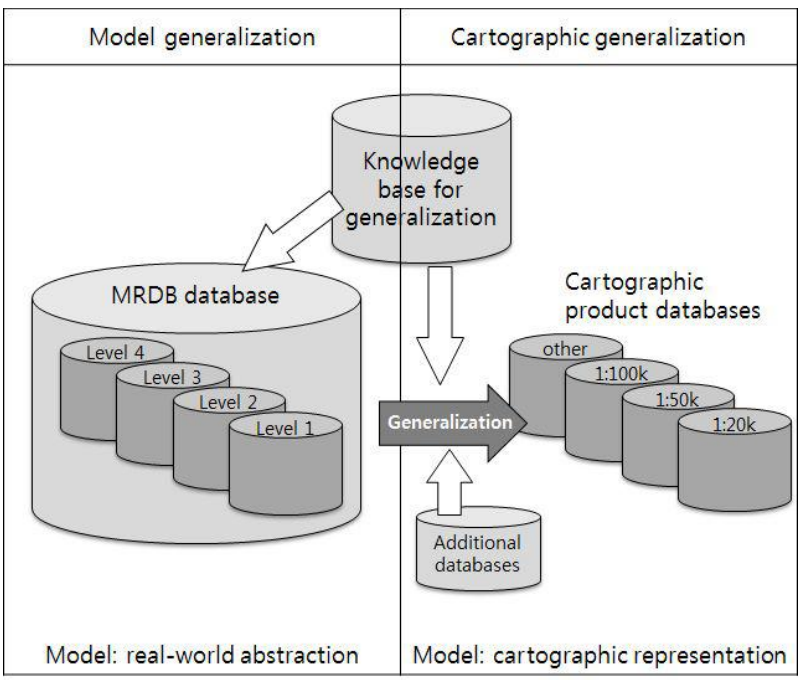
1.1. 연구 배경 및 목적

지금까지 공간정보 구축은 주로 수작업에 의존하여 정보의 취득, 가공 등의 과정이 수행되어왔다. 따라서 공간 데이터베이스의 구축을 위해 큰 비용과 시간이 요구되고 있는 현실이다. 이에 따라 이러한 데이터 구축 과정을 중복 없이 더 효율적으로 수행하기 위한 여러 가지 방법에 관한 연구가 진행됐다(최신영 등, 1998; Li, 2007).

그중에서 데이터 구축 분야에서 효율적이라고 알려진 방법은 기존에 구축된 대축척의 데이터를 지도 일반화 기법을 적용하여 소축척의 데이터로 바꾸는 방법이다. 지도 일반화는 축척 또는 제작자의 필요나 관점에 따라 다르게 적용될 수 있으며 대축척 지도(원도)의 지형지물 표현에 관련된 다양한 의사결정 과정을 포함한다. 즉 지도 일반화는 지형지물 중 일반화의 대상이 될 요소들을 선택하여 축척에 따라 추상화하는 원리와 방법이라고 할 수 있다(한균형, 1996; 이민부 외, 2001). 특히 수치지형도(digital map)에서 일반화는 대축척에서 소축척으로의 일방적인 흐름에 따르며, 복잡성을 감소시키고, 공간 및 속성 정확도를 유지하며, 자료의 미적 품질과 논리적 체계를 유지함은 물론 일반화를 위한 규칙을 적용하여야 하는 등 이론적 기본 요건을 만족하여야 한다(Ruas and Plazanet, 1996).

지도 일반화의 과정은 개념적으로 크게 두 가지로 나누어 볼 수 있다. 하나는 모델 일반화 과정이고 다른 하나는 지도학적 일반화 과정이다. 모델 일반화 과정은 원 지도 데이터로부터 압축된 지도 데이터를 추출하

는 데 초점을 두고 있으며, 객체의 구조에 기반을 둔 필터링을 통해 지도 객체를 제거하거나 병합하는 과정을 포함한다. 지도학적 일반화 기법은 지도상의 객체들을 축척에 맞게 표현할 때 불필요한 상세함을 제거하거나 의도적으로 강조하는 데에 초점을 두고 있다. 두 과정은 상호 보완적인 관계에 있으며 보통 모델 일반화 과정은 지도학적 일반화 과정을 위한 예비 단계로 사용된다(Weibel and Jones, 1998). <그림 1-1>은 모델 일반화 과정과 지도학적 일반화 과정의 개념을 표현한 것이다.



<그림 1-1> 모델 일반화와 지도학적 일반화(Kilpelainen, 2000)

모델 일반화와 지도학적 일반화 과정에서 지도 객체의 변화에 작용하는 요소들을 일반화 연산자(operator)라고 지칭한다. Li(2006)의 연구에서는 지도를 구성하는 객체들의 기하학적 특징, 즉 점, 선 및 면을 기반으

로 필수적인 일반화 연산자들을 분류하였다(<표 1-1>). 모델 일반화 과정에서는 삭제(elimination), 병합(aggregation 또는 amalgamation) 등의 연산자가 주로 사용되며, 지도학적 일반화 과정에서는 단순화(simplification), 완만화(smoothing), 과장(exaggeration)들의 연산자를 주로 사용한다(Müller *et al.*, 1995).

<표 1-1> 지도 일반화를 위한 필수 연산자들(Li, 2006)

객체 요소	주 연산자	부 연산자
점 요소(개별)	·이동 ·삭제 ·확대	
점 요소(군집)	·확대 ·지역화(regionalization) ·단순화 ·상징화(typification)	
선 요소(개별)	·이동 ·삭제 ·부분변환(partial modification) ·완만화(smoothing)	·curve-fitting ·필터링
선 요소(군집)	·이동 ·병합 ·삭제 ·분리	·double-to-singleline ·ringing-to-point
면 요소(개별)	·분리 ·이동 ·삭제 ·확대 ·단순화 ·분할	·area-to-point ·area-to-line ·directional thickening ·enlargement
면 요소(군집)	·확대 ·융합 ·병합 ·단순화 ·상징화 ·삭제	

지도 일반화에 관한 연구는 1960년대부터 꾸준히 진행되어 오고 있다(Li, 2007; Zhou, 2014; 김남신, 2006). 초기의 지도 일반화는 대부분의 과정을 지도 제작 전문가의 수작업에 의존해 왔다. 그런데 수작업에 의한 일반화는 효율성이 떨어지고 일관성이 모자라는 문제점이 있다. 1960년대 이래로 이러한 문제를 보완하기 위해 수학적 원리를 적용한 다양한 일반화 기법들이 개발되었다(Perkal, 1966; Cromely, 1991; Müller *et al.,* 1995). 이 기법들은 주로 지도 객체의 기하학적인 형태 변형을 위한 알고리듬 개발에 중점을 두었기 때문에, 해안선이나 등고선과 같은 선형 요소를 단순화시키는 알고리듬 위주로 개발되었으며, 점이나 면 요소에 대한 일반화는 상대적으로 연구가 활발하게 진행되지 못하였다.

1980년대 이후에는 지도를 구성하는 각각의 객체들에 맞게 개발되던 알고리듬 연구의 한계를 극복하고 다양한 지도요소들에 대해 복합적으로 적용 가능한 일반화 방법의 개발 필요성이 대두되었다. 이에 따라 일반화의 대상이 되는 객체와 적용할 연산자들을 하나의 모델로 구성하여 지도 일반화를 수행하는 모델링에 의한 일반화 연구들이 활발하게 진행되었다(McMaster, 1991; Müller *et al.*, 1995; Buttenfield, 1991; 박환철, 2000; 이민파, 2001). 모델링에 의한 지도 일반화는 GIS(Geographic Information System)의 발달과 더불어 지도 일반화에 많은 변화를 가져왔고, 이를 바탕으로 숙련된 지도 제작자들이 지도를 제작할 때 사용하는 지식과 지리학적 원리를 일반화에 적용하려는 시도인 규칙기반(rule-based) 일반화 기법에 관한 연구가 이루어지게 되었다.

규칙기반 일반화에 사용되는 지도학적 원리와 지식은 크게 세 가지 방법을 통해서 얻을 수 있다. 첫째, 실제 지도를 제작하는 지도 제작자들과의 인터뷰로부터 얻는 방법이다. 이 방법은 주로 지도 제작자들을 인터뷰하고 그 결과를 토대로 그들의 작업 과정을 문서화 하는 방식으로

진행된다. 둘째, 기존의 지도를 분석함으로써 얻는 방법이다. 이 방법은 같은 지역의 대축척 지도와 소축척 지도 간의 변화 분석을 통해 적용된 지도학적 원칙을 발견한다. 그리고 마지막으로 지도 제작 시 지켜야 할 제작 규격(specification)을 통해 얻는 방법이다(Li, 2007). 이러한 지도학적 원리나 지식의 습득 방법에 관한 연구는 규칙기반 일반화가 등장한 이래로 꾸준히 진행되었다(Buttenfield and McMaster, 1991; Weibel, 1995).

지도 제작자와의 인터뷰로부터 지식을 추출하는 방법은 그 과정과 도출되는 지식이 명확하지 않다는 단점이 있다. 또한, 완성된 지도를 분석하여 지도 제작에 대한 지식을 얻고자 하는 두 번째 방법 역시 일부 지도 일반화 연산자에 관한 제한적 연구만이 있으며, 점차 실험 대상 객체와 면적 등을 넓혀 나가고 있지만, 아직 확실한 결과를 내고 있다고 보기는 어렵다(Töpfer and Pillewizer, 1966; Leitner and Buttenfield, 1995; Li and Choi, 2002). 마지막으로, 지도의 사양에서 지식을 습득하는 방법 역시 어려움이 있다. 그 이유는 지도의 사양에는 일반적으로 지도 제작에 있어서 반드시 수행되어야 할 과업목록보다는 수행되지 말아야 할 과업목록에 관한 내용이 대부분이기 때문이다. 따라서 일반화를 자동으로 수행하기에 충분한 규칙을 도출하는 것은 많은 어려움이 있는 것으로 알려져 있다(Li, 2007).

규칙기반 일반화에 사용될 규칙을 도출하고자 하는 연구들이 활발하게 진행되는 한편 실제로 지도를 제작하는 국가 지도 제작기관(National Map Agency, NMA)들은 규칙기반 일반화의 한계를 극복하고 보다 효율적인 지도 일반화를 수행하기 위해 대용량의 지도 데이터에 대해 일괄적으로 적용할 수 있는 일반화 방법에 관한 연구를 진행하고 있다(Neun *et al.,* 2004; Lamy *et al.*, 2002; Galanda and Weibel, 2002; Gaffuri,

2006; Burghardt *et al.,* 2014). 특히 네덜란드의 지도 제작기관인 Dutch Kadaster에서는 네덜란드의 1:10,000 축척 지도 데이터베이스인 Top10NL을 관리하던 기존 방식이 비효율적이고 비용이 많이 소요되기 때문에 새로운 방법을 개발하려는 다양한 시도를 하였다. 그들은 1:50,000 지도를 제작하는 과정에서 다양한 객체들의 변화에 대해 필요한 지식을 지도 제작자들로부터 얻어내는 한편 1:50,000 데이터와 1:10,000 데이터의 비교 분석을 통해 1:10,000 지도에서 1:50,000 축척의 지도로 변환될 때 사용된 중요한 규칙들을 도출하였다. 그 이후 도출한 정보와 상용 GIS 소프트웨어의 기능들을 통합하여 지도 제작자의 지식을 반영한 정교한 모델을 만들고 다른 축척의 지도 제작을 완전히 자동화하는 구성을 시도하였다(Stoter *et al.*, 2009; Stoter *et al.,* 2014).

우리나라의 지도 제작기관인 국토지리정보원에서도 1:5,000 수치지형도를 1:25,000 수치지형도로 변환하는 축소 편집 과정의 자동화를 위해 연구 공개설명회를 개최하고 보고서를 발간하는 등(국토지리정보원, 2012, 2014) 다각도의 노력을 기울이고 있다. 다만 현재까지 실무에 적용할 만한 구체적인 성과가 없어서 실무 현장에서는 숙련된 편집자에 의존하여 축소 편집 과정이 진행되고 있다.

현재까지의 연구 및 실무 적용의 측면에서 봤을 때, 지도 일반화의 방향은 최대한 사람의 개입을 배제한 채 지도의 품질이 담보되는 자동 일반화를 수행하는 방향으로 진행되고 있다고 할 수 있다. 이처럼 지도 일반화 과정에서 사람의 개입을 배제하는 방향의 연구가 꾸준히 진행되고 있지만, 실제 지도 일반화 과정에서 사람의 영향이 얼마나 있는지, 지도 일반화 결과에 어떠한 차이를 가져오는지에 대한 연구는 상대적으로 부족한 실정이다. 사람의 개입 영향에 관한 연구가 존재하긴 하지만 지도 제작 과정에서 자료 취득 시 발생하는 오류에 대한 내용이거나(Biljecki

et al., 2018) 다른 시간에 제작된 지도간의 표현 차이를 비교하는 연구(Schaffer *et al.*, 2016)가 대부분이다. 지도 일반화에 있어서 편집자마다 숙련도나 작업 과정 등에서 차이가 있다는 사실은 기존 연구 및 현업에서 지속적인 문제가 제기되어왔던 부분이다. 그러나 축소 편집과 같은 지도 일반화 분야에서 제작자 개인이 일반화 과정에 미치는 영향에 대해서 정량적으로 측정하거나, 제작자 간 편차가 어떤 형태로 존재하는지에 대한 연구는 미진한 편이다. 지도 데이터 품질의 균질성 저하 문제의 양상을 파악하고, 이를 통해 축소 편집 등 일반화에 사용되는 규정의 미진한 부분을 보완하기 위해서는 축소 편집 시 지도 제작자 간의 편차가 어느 정도 있는지, 어떠한 양상을 보이는지에 대해 정확하게 분석하는 과정이 필요하다.

본 연구에서는 기계학습 기법을 활용하여 지도 제작자 간의 차이가 지도 일반화 결과에 어떠한 영향을 미치는지 정량적으로 분석하고 나아가 축소 편집 규정의 보완 방안을 도출하고자 하였다. 분석 대상 객체로는 건물과 도로객체를 대상으로 지도 일반화 과정, 특히 선택적 삭제 과정에서 사람의 개입으로 인해 발생하는 오차를 정량화하고자 하였다. 해당 객체들을 대상으로 기계학습 기법을 활용하여 기계학습으로부터 생성된 모델이 서로 다른 지도 제작자들의 편집 결과물들에서 어떤 정확도를 나타내는지를 비교 분석하고자 하였다. 기계학습 기법을 활용함으로써 얻을 수 있는 장점은 서로 다른 제작자가 편집한 지도 일반화 결과물을 일관적인 기준으로 평가할 수 있다는 점이다. 또한, 도시와 비 도시 지역 각각에 대해 기계학습 모델을 생성하고 그 결과를 분석함으로써 건물과 도로객체의 선택적 삭제에 영향을 주는 주요 속성들을 파악하고, 지역의 특성에 알맞은 기계학습 알고리듬 임곗값을 설정하여 이를 관련 규정의 보완에 활용하고자 하였다.

기계학습의 적용을 위해서는 먼저 지도 일반화를 기계학습이 적용 가능한 문제로 정의하고 접근할 필요가 있다. 본 연구에서는 지도 일반화 중에서도 객체의 선택적 삭제를 지도 학습(supervised learning)으로 해결 가능한 문제로 정의하고 기계학습 기법을 적용하기로 하였다. 지도학습은 기계학습의 한 종류로, 데이터를 입력할 때 입력 데이터의 속성과 그 데이터의 정답(label)을 함께 입력하여 학습시키는 방법이다. 예측하려는 값의 성질에 따라 크게 분류문제와 회귀 문제로 나뉜다.

본 연구에서는 지도 제작자의 축소 편집 결과물에서의 객체 선택적 삭제 여부를 정답으로 입력받아 분류 과정을 수행하는 문제로 간주하고 기계학습 기법을 적용하였다. 기계학습 기법의 적용 과정은 다음과 같다. 첫째, 각 객체의 삭제 여부를 결정하기 위한 속성(건물의 경우 면적 등, 도로의 경우 길이 등)이 입력 속성으로 사용되고 각 객체가 소축척에서 삭제되었는지가 출력 속성으로 사용된다. 이렇게 입력 및 출력 속성을 가지고 있는 데이터를 통해 기계학습 모델을 학습시키게 된다. 마지막으로, 학습된 모델을 다른 지역, 즉 입력 속성은 가지고 있지만, 출력 속성(삭제 여부)은 가지고 있지 않은 데이터에 적용하여 해당 지역의 객체들이 소축척으로 일반화되었을 때 존치 여부를 예측하는 것이다.

본 연구에서는 기계학습 알고리듬 중에 객체의 선택적 삭제에 적용 가능한 의사결정 나무(decision tree), k-최근접 이웃(k-nearest neighbor), SVM(Support Vector Machine) 및 인공신경망(Artificial Neural Network, ANN)과 같은 4가지의 알고리듬을 채택하여 1:5,000 수치지형도에서 1:25,000 수치지형도로의 축소 편집 시 건물과 도로객체의 삭제 여부를 예측할 수 있는 기계학습 모델을 생성하였다. 첫 번째로는 생성된 모델을 서로 다른 6명의 지도 제작자의 결과물에 적용하여 객체의 삭제 여부를 예측한 예측률을 비교하였다. 예측 정확도의 비교는 전

체적인 예측률 및 건물과 도로의 특성에 따른 예측률에 대해서 이루어졌으며, 이를 통해 지도 편집자 간의 편집 기법의 차이가 존재함을 보이고자 하였다.

두 번째로는 도시지역과 비도시 지역 각각에 대해 해당 지역에 최적화된 기계학습 모델을 생성하고 그 결과를 분석함으로써 각 지역에서 건물과 도로의 선택적 삭제에 주된 영향을 미치는 속성들을 분석하고자 하였다. 나아가 분석된 속성값들을 지도 일반화에 적용함으로써 객체의 선택적 삭제에 있어서 현재 모호하게 제정된 축소 편집 관련 규정의 개정안을 도출하고자 하였다.

1.2. 연구 동향

지도 일반화에 관련한 연구들은 다양한 방면으로 연구되고 있다. 그 과정에서 앞서 언급한 지도 일반화 연산자들이 개발되었고, 더욱 정확한 일반화 결과를 얻기 위해 다양한 연산자들을 조합하여 활용하는 연구들이 진행되어왔다. 또한, 지도 제작에 활용되는 규칙들을 유형화하여 일반화를 자동으로 수행하려는 다수의 연구가 진행됐다. 본 장에서는 지도 일반화 기법에 관한 연구들을 살펴보고, 추가로 지도 일반화 분야의 문제 해결을 위해 기계학습 기법을 도입하고자 하는 최신의 연구들을 검토하였다. 마지막으로, 현재의 지도 제작 및 소축척 지도 제작 현황을 알아보고 본 연구가 가지는 의미와 앞으로의 적용 가능성을 검토하였다.

1.2.1. 건물과 도로의 일반화 기법

건물과 도로라는 지도를 구성하는 가장 중요한 객체들을 대상으로 한 일반화 연구는 오랜 시간 꾸준히 진행됐다. 그 과정의 대부분은 지도 제작 전문가들의 수작업에 의존해 온 노동집약적인 과정이었다. 그러나 이러한 수작업에 의한 지도 일반화는 비용과 시간 면에서 효율적이라고 볼 수 없을 뿐 아니라, 사람이 진행하는 작업이라는 특성상 그 결과물의 일관성을 보장하기 힘들다는 문제가 있다. 초창기의 지도 일반화 연구들은 이러한 문제를 수학적·기하학적 원리를 적용하여 해결하려고 시도하였다.

수학적, 기하학적 원리를 적용하려는 시도는 먼저 지도의 선형 객체들에 관한 연구로 시작되었다. 대표적으로 가장 널리 사용되는 선형 단순화 알고리듬인 Douglas-Peucker 알고리듬을 제안한 Douglas and

Peucker(1973)의 연구가 있으며, 그 외에도 Sleeve-fitting 알고리듬(Zhao and Saalfeld, 1997), Lang 알고리듬(Lang, 1969), Reumann-Witkam 알고리듬(Reumann and Witkam, 1974), Visvalingam-Whyatt 알고리듬(Visvalingam and Whyatt, 1993), 선회 각 함수(turning function)를 이용한 단순화(Rangayyan *et al.,* 2008) 등 다양한 알고리듬들의 개발이 진행됐다.

1980년대에 이르러서는 선형 객체뿐 아니라 건물이나 주거지 등 면형 객체에 대한 기하학적 일반화 연구들이 활발하게 진행되었다. 기하학적 일반화 연구들(Li *et al.,* 1995; Su *et al.,* 1997; 1998)에서는 수학적 형태학을 기반으로 면 객체의 병합, 면 객체의 삭제, 면 객체의 선형 단순화, 부분적 삭제, 소규모 지역 일반화 등 다양한 일반화 연산자들을 적용하려고 시도하였다.

최적화 기법 등 다른 분야에서 사용되는 기법들을 기하학적 일반화에 적용하려는 연구들도 있었다. Regnauld(1996) 그래프 이론에서 사용되는 Minimum Spanning Tree(MST)라는 기법을 사용하여 건물을 군집화하고자 하였다. 그의 연구에서는 건물의 개수에 따라 2가지 형태로 군집을 생성하였고, 시각적(정성적) 평가를 시도하였지만, 도시지역 일부에 대한 실험에 그쳤다. Anders and Sester(2000) 는 계층적 클러스터 알고리듬을 사용하여 비슷한 특성을 가지는 건물들과 호수를 군집화하려고 시도하였다. 그들의 연구에서는 벡터와 래스터 자료를 모두 활용하여 사용자가 입력변수를 조절할 필요 없이 건물의 군집을 찾아낼 수 있었음을 보였으나, 대상 지역이 도시의 극히 일부에 한정되는 한편 데이터의 전처리에 많은 시간이 소요되는 단점이 있었다. Ware and Jones(1998)의 연구에서는 최적화 기법의 하나인 담금질 기법(Simulated Annealing, SA)을 활용해서 건물의 이동(displacement)을 효율적으로 수행하려는 시도

가 있었다. 그들의 연구는 결과적으로 건물끼리의 충돌 없이 성공적인 건물의 이동이 가능함을 보였으나, 5개 지역의 총 50여 개의 건물 데이터만 검증됨을 보였다.

이후 더 넓은 지역에 대해 효율적으로 일반화를 수행하기 위해 규칙기반(rule-based) 일반화에 관한 연구들이 활발하게 진행되었다. 규칙기반 일반화를 위해서는 일반화에 사용될 규칙들, 즉 그동안 소축척 지도 제작에 사용되어왔던 지식을 습득하는 것이 관건이 되었다. 규칙기반 일반화에 사용되는 지식은 지도 제작자와의 인터뷰, 기존 지도 데이터의 분석, 또는 지도 제작 규정 등을 통해 얻을 수 있다고 알려져 있다(Li, 2007). 그러나 지도 제작 규정이나 지도 제작자와의 인터뷰를 통해서 얻을 수 있는 지식에는 한계가 있다고 알려져 있으며(Müller *et al.,* 1995), 이미 구축된 지도를 분석하여 지식을 습득하려는 연구들이 주로 진행됐다.

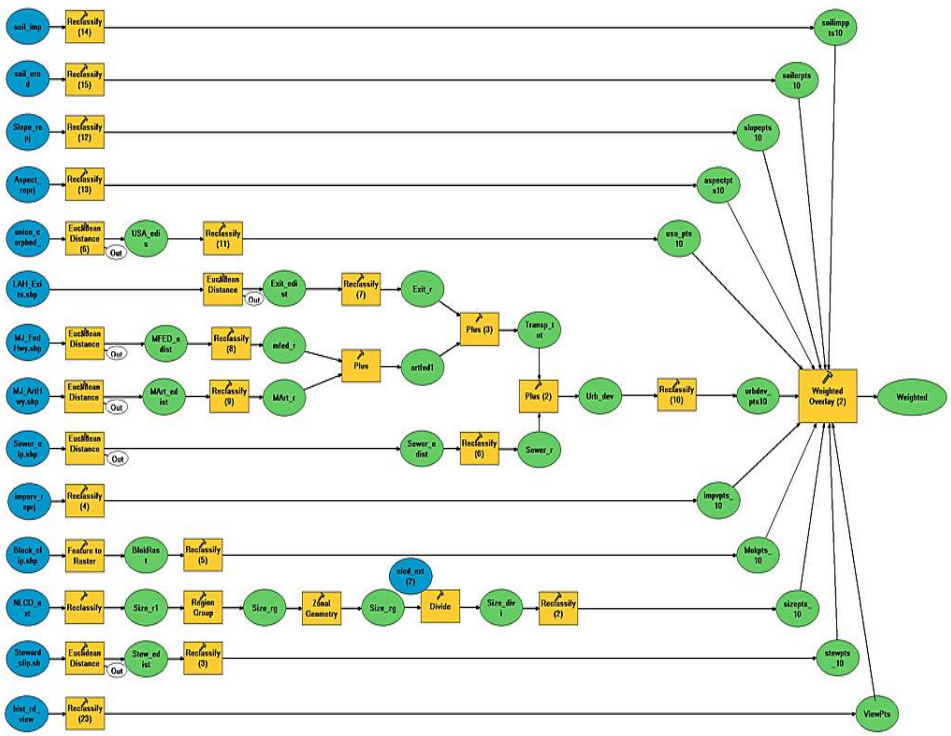
구축된 지도를 분석하려는 대표적인 연구로는 Töpfer(1966)의 연구가 있다. 이 연구에서는 서로 다른 축척의 지도 객체들을 분석하여 달라지는 축척에 따라 남겨져야 할 건물의 개수를 결정하는 방정식을 정립하였다. 이 방정식은 “Töpfer의 radical law”로 널리 알려져 있다. 이 방정식은 최근의 연구들에서도 많이 활용되고 있는데, 박우진 등(2013)의 연구에서는 Töpfer의 radical law와 로짓 모델을 활용하여 도로 네트워크에 상대적인 중요도를 부여하고, 상대적으로 더 중요한 도로들을 선택하여 남기는 방법을 제안하였다. Radical law를 활용한 다른 연구들도 다수 존재하지만(Downs and Mackaness, 2002; Wilmer and Brewer, 2010; Jiang *et al.,* 2013), 남겨져야 할 객체의 개수에 대한 기준만 제시했을 뿐 어떠한 객체가 남겨지고 어떠한 객체가 삭제될지에 대한 근거를 뚜렷이 제시했다고 보기엔 어렵다. 국가지도제작기관 주도로 규칙기반 일반

화를 시도한 연구 중 Stoter *et al.*(2014)의 연구에서는 네덜란드의 1:10,000 지도로부터 1:50,000 지도를 자동으로 생성하려고 하였다. 이 연구에서는 네덜란드 전역의 지도를 대상으로 자동 일반화를 시도하였는데, 정성적 및 정량적 평가에서 모두 만족할만한 정확도를 보였다. 그러나 일반화 과정을 적용하기 위한 데이터 전처리 과정에서 불가피한 사람의 개입이 발생했으며, 이 과정에서 많은 시간과 비용이 필요하였다. 또한, 지도 제작자들의 지도 제작 방법을 규칙화하는 데에도 어려움을 겪었던 것을 볼 수 있었다.

이 외에도 복수의 지도 일반화 연산자들을 조합하여 일반화 과정을 자동화하려는 연구들도 꾸준히 진행됐다. Li *et al.*(2004)의 연구에서는 도시 형태학과 게슈탈트 이론을 접목해서 건물들을 그룹화하고, 각각의 건물 그룹에 적절한 연산자를 선정하여 건물 일반화를 시도하였고, 4개의 서로 다른 축척에 대해서 만족할 만한 정확도를 보였다. 하지만 실험 대상 건물이 총 87개에 불과하며, 자신도 지적하듯이 다양한 건물 패턴에서 좋은 결과를 나타낼지 장담할 수 없다는 문제가 있다. Wang and Doihara(2004)의 연구에서는 도로와 건물을 함께 고려한 일반화를 수행하였다. 이들은 도로에 대한 일반화를 먼저 시도한 후, 도로 일반화의 결과에 따라 다시 건물을 그룹화하여 일반화하는 순차적인 방법을 시도하였다. 먼저 도로 폴리곤을 도로 네트워크의 형태로 변환시키고, 건물 일반화에 대해서는 병합만을 수행하였다. 그 결과 생성된 서로 다른 3가지의 모듈을 적용하여 정성적 평가를 진행하였고 만족스러운 성과를 보였으나 입력 데이터에 막힌 도로가 있거나 교차로의 경우에는 정확도가 낮게 나타나는 현상들이 있었다.

Damen *et al.*(2007)의 연구에서는 형태학적 연산자를 채택하여 건물의 병합을 효율적으로 수행하고자 하였다. 그들의 연구에서는 건물과 건물

간의 형태를 열린 부분과 닫힌 부분으로 구분하고 이를 형태학적으로 분석하여 건물 병합을 효율적으로 하고자 시도하였고, 나아가 선형 단순화도 형태학적 요소들을 적용하고자 하였다. 최근 연구의 경우 Vetter *et al.*(2015)의 연구에서 스위스의 Topographic Landscape Model(TLM) 데이터의 개별 주택 건물에 대한 자동 일반화를 시도하였다. 그들은 <그림 1-2>와 같이 ArcGIS 소프트웨어에서 제공하는 다양한 지도 일반화 연산자들을 ModelBuilder를 활용하여 조합하는 것으로 자동 일반화를 수행하고자 하였다. 그 결과 전반적으로 좋은 정확도를 보였으나, 건물의 형태에 따라서 일정하지 않은 결과를 보였으며, 또한 구시가지와 같은 구조에서는 낮은 정확도를 보이는 문제가 있었다.



<그림 1-2> Vetter *et al.*(2015)의 연구에 활용된 ArcGIS ModelBuilder

일반화 연산자들을 조합하여 자체적으로 소프트웨어를 제작한 사례도 있다. Modiri *et al.*(2014)는 비정부단체에서 제공하는 데이터의 지형지물에 대해서 점, 선, 면 객체에 따라 일반화를 수행하는 소프트웨어를 개발하였다. 그러나 개발 결과에 대한 정확도 평가가 수행되지 않아 소프트웨어의 결과물을 객관적으로 평가할 수 없는 문제가 있다.

건물 일반화에 관한 최근의 연구들은 건물의 군집화(clustering)에 관한 연구들이 주류를 이루고 있다. Deng *et al.*(2018)의 연구에서는 게슈탈트 이론을 활용한 9개의 군집화 방법을 건물에 적용하여 각각의 알고리듬에 대한 평가를 수행하였고, 각각의 알고리듬에 대한 장·단점을 분석하였다. Wei *et al.*(2018)은 건물의 군집화에 영향을 주는 특성들을 크기(size), 모양(shape), 방향(orientation)으로 구별한 뒤, 중요한 특성들을 주성분 분석(Principal Component Analysis, PCA)을 통해 분석하였다. 그러나 건물의 높이나 건물 간의 관계 등 다른 특징들을 포함하지 못하였고, 또한 건물의 특성을 측정할 때 복잡한 매개 변수의 결정 과정이 별도로 필요하였다. Pilehforooshha and Karimi(2019)는 건물들의 선형적 패턴을 분석하여 건물군 단위로 군집화를 시도하였다. 제안된 기법을 통해 기존의 임곗값 설정 과정을 단순화하고, 만족할만한 일반화 결과를 도출해 내는 데 성공하였다. 그러나 패턴 분석을 위한 임곗값의 설정 과정이 필요하며 선형적 패턴 이외에 다른 요소들은 고려되지 않았다는 문제점이 있다.

건물과는 다르게 도로 일반화에 관한 최근의 연구들은 선택적 삭제에 관한 연구들이 활발하게 진행되고 있다. Chen *et al.*(2009)는 mesh density를 이용하여 도로의 선택적 삭제를 수행하는 기법을 제안하였다. 이를 위해 먼저 임곗값을 주어 mesh의 후보군을 추출하고, 추출된 후보군에 대해 mesh의 밀도를 측정하여 선택적으로 도로를 삭제하였다. 이

연구는 도로의 지역적인 특징(도시부, 교외 지역 등)에 따라 정확도가 78%에서 96%까지 큰 편차를 보이는 문제가 있었다. Shoman and Gülgen(2017)은 계층 구조를 활용하여 도로 네트워크의 일반화를 수행하였다. 이들은 도로의 기능적 요소와 형상적 요소를 고려하여 계층 구조를 생성하고, 생성된 계층 구조에 따른 순차적인 일반화를 시도하였고 만족할만한 성과를 보였다. 그러나 계층 구조를 생성하는 과정에 있어서 사용자가 매개 변수를 직접 설정해야 하는 과정이 복잡하다는 문제가 있다.

앞서 언급한 연구들을 종합하여 검토한 결과, 기존의 연구들에서는 공통으로 일반화 과정에서 사람의 개입이 필요하며, 또한 데이터 품질의 감소 등 이로부터 야기되는 문제가 발생한다는 점을 지적하고 있다. 그러나 사람의 개입으로 인해 발생한 오차들이 어느 정도이며 어떤 양상을 보이는지를 분석한 연구는 이뤄지지 않고 있다.

1.2.2. 기계학습 기법을 적용한 연구

기계학습 기법의 적용은 GIS 분야에서도 활발하게 이루어지고 있는데, 비단 일반화 부분만이 아니라 다양한 부분에서도 이러한 시도들을 찾아볼 수 있었다. 초창기의 시도는 Lagrange *et al.*(2000)의 연구에서 찾아볼 수 있다. 이 연구에서는 기존의 지도 일반화 연구에서 알고리듬들을 도출해 내는 과정이 복잡하고 유형화시키기 어렵다는 것을 지적하며, 기계학습 기법 도입의 필요성을 강조하였다. 기계학습 기법이 지도 일반화 알고리듬의 자동화에 이바지할 수 있을 것이라는 분석을 하였지만, 실제로 기계학습 기법을 적용한 실험은 이루어지지 못하였다. Balboa and Lopez(2008)는 인공신경망을 사용한 도로 분류 기법을 제안

하였다. 그들은 도로 데이터로부터 주성분 분석을 활용하여 도로를 분류할 수 있는 값들을 추출한 후, 도로를 총 5가지의 범주(very smooth, smooth, sinuous with stable directionality, sinuous with variable directionality, and very sinuous)로 분류하였으며 이 범주들은 출력 클래스로 사용되었다. 그들의 연구에서는 14개 유닛의 입력층, 3개 유닛의 은닉층, 5개 유닛의 출력층으로 이루어진 총 3개의 층을 가지는 Back-Propagation Neural Network(BPNN)를 사용하였고, cross validation을 통해 도로 분류 결과의 정확도를 평가하였다. 이들의 연구에서 인공신경망은 지도의 일반화가 아니라 도로 데이터의 분류만을 위해서 사용되었지만, 도로의 분류 결과가 도로의 중요도를 판단할 수 있는 근거가 되며 또한 후에 객체 선택의 기준으로 사용될 수 있으므로 일반화 연구와도 연관성이 있다고 할 수 있다.

지도 일반화 분야에 기계학습 기법을 적용하기 시작한 시도들은 2010년대 중반부터 활발하게 이루어지기 시작했다. Zhou and Li(2014)의 연구에서는 소축척 지도의 도로 네트워크 갱신을 위해 Self Organizing Map(SOM)과 BPNN을 적용한 기법을 제안하였다. 여기에서 SOM은 비지도 학습, 클러스터링 방식으로 도로 네트워크를 선택하고, BPNN은 지도 학습의 분류 방식으로 도로 네트워크를 선택해서, 각각의 방법들의 정확도를 평가하고, 실제 데이터와의 비교를 통해 정확도를 평가하였다. Karsznia and Weibel(2017)의 연구에서는 데이터 응축과 기계학습 방법을 사용하여 소축척 지도에서의 거주지 선택을 향상하려 했다. 그들은 앞선 연구에서 들과 달리 기계학습을 사용하여 새로운 지도 일반화 규칙 도출을 시도하였다. 이를 위해 먼저 성분 분석과 지도 데이터의 분석을 통해 거주지의 유지 및 삭제에 영향을 주는 요인을 분석하고, 의사결정나무 알고리듬을 적용하여 기존의 지도 일반화 규칙과는 차별화되는 규

칙을 도출하고자 하였다. 이 연구는 본 연구에서 시도하고자 하는 방법과 유사한 점들이 많지만, 적용 대상 데이터가 면형 데이터에 국한되고 의사결정 나무 외에 다른 기계학습 기법을 적용하지 않고 의사결정 나무에 의한 분류 결과의 분석에 집중한 점에서 본 연구에서 사용한 방법과는 차이가 있다.

원격 탐사 분야에서도 기계학습, 나아가 딥러닝 기법을 지도 데이터에 적용하려는 시도들도 있었다. Alshehhi *et al.*(2017)의 연구에서는 CNN(Convolutional Neural Network)를 활용하여 1m 해상도의 위성 영상에서부터 도시지역의 도로와 건물 객체를 추출하는 방법을 제안하였다. 다른 기법들과의 비교 평가에서 CNN을 활용한 제안된 기법이 우수함을 보였으나, 영상의 가장자리 부분에서는 추가적인 처리가 필요하다는 문제점이 있었다. Nogueira *et al.*(2017)의 연구에서도 마찬가지로 CNN을 활용하여 경관을 분류하는 시도가 있었다. 이 연구에서는 서로 다른 6개의 CNN을 구성하여 각각의 3가지의 서로 다른 지역에 대해서 정확도 평가를 수행하였고, 만족할만한 결과를 나타냈다. 이 외에도 사용자들이 코드를 공유하는 가장 큰 사이트인 GitHub에서도 딥러닝 기법으로 지도를 자동 생성하려는 DeepOSM 프로젝트를 진행하기도 하였다. DeepOSM은 미국 델라웨어 지역의 200㎢ 이상의 면적에 대해 Fully-connected neural network를 활용하여 도로객체를 추출하려고 시도하였고, 수 분 이내로 모든 과정이 완료되었으며, 전체적으로 75~80% 사이의 분류 정확도를 보였다. DeepOSM의 결과로 추출된 도로객체는 <그림 1-3>과 같다. 그러나 더는 정확도를 높이지 못하고 현재는 프로젝트가 잠정적으로 중단된 상태이다.

<그림 1-3> DeepOSM의 결과로 추출된 도로(굵은 선)

1.2.3. 지도 제작 및 일반화 과정

우리나라의 경우 국가기본도의 활용성과 최신성을 위하여 2002년 이후 5년 주기 갱신체계를 도입하여 2006년까지는 전국을 5권역으로 하여 진행하였으며, 2007년도부터는 갱신 단위를 전국 4권역과 광역시로 개편하고 2년 주기로 갱신을 시행하였다. 그리고 2011년부터는 전국을 2권역으로 하는 갱신체계와 수시갱신 체계를 도입하여 국가기본도를 수정·갱신하였으며, 현재는 전국을 2권역으로 하여 한 권역은 정시수정, 다른 한 권역은 수시갱신 체계를 적용하여 전국을 대상으로 갱신을 하고 있다. 수시수정 방식의 경우 준공도면 등의 자료를 수집하고 관찰하여 주 단위로 갱신이 이루어지고 있다. 수시수정의 대상은 국민 생활과 밀접하고 변화가 많은 지형지물을 중심으로 변화정보를 즉시 지도에 반영하게 되어있으며, 주요 건물 55종(아파트, 관공서, 병원, 대형할인점 등), 도로나 교량 등의 신설, 또는 택지개발 등이 그 대상이 된다(<그림 1-4>).

<그림 1-4> 수치지형도 수시수정 주요 내용
(서울시 GIS(gis.seoul.go.kr))

현재 국토지리정보원에서는 다양한 축척의 수치지형도를 제작하고 관리하고 있다. 현재 우리나라에서 제작되고 있는 수치지형도는 축척에 따라 1:1,000, 1:2,500, 1:5,000, 1:25,000, 1:250,000 등이 있다. 이 중 1:5,000 수치지형도와 1:25,000 수치지형도는 국토지리정보원에서 전 국토에 대

해 제작, 관리하고 있으며, 1:1,000 수치지형도의 경우 주로 시가지 지역을 대상으로 지방자치단체별로 제작되고 있다.

수치지형도 2.0에서 표현하고 있는 지형지물은 교통, 수계 및 해양, 건물 및 시설물, 문화 및 시설, 식생, 지형, 주기 등 총 8개 대분류와 104개의 소분류로 나누어 구성되어 있다. 지형지물 유일식별자(Unique Feature IDentifier, UFID)를 가지고 있어 공간 데이터와 속성 데이터를 결합하여 사용할 수 있으며 단순한 도형 정보가 아닌 위상정보(topology)를 포함할 수 있도록 개선되었다. 또한, 수치지형도 2.0부터는 DXF 포맷뿐만 아니라 국토지리정보원 내부 포맷인 NGI 포맷으로도 제작, 판매되고 있다. 기존의 수치지형도 1.0이 기본지리정보 구축 용도로 사용되기 위해서는 별도의 복잡한 작업공정의 필요했지만 2.0 버전에서는 손쉽게 기본지리정보로 활용될 수 있게 제작되었다(<표 1-2>).

<표 1-2> 수치지형도 1.0과 수치지형도 2.0

구분	수치지형도 1.0	수치지형도 2.0
분류체계	축척에 따라 서로 다른 코드체계 사용 597개의 소분류로 구성	축척과 관계없이 같은 코드 분류체계 사용 8개 대분류, 104개 소분류로 구성
단위	도엽단계 파일	도엽단계 파일
UFID	없음	있음
속성항목	없음	다양한 속성항목
위상정보 표현	없음	위상정보 구축 가능
데이터 구조	도형구조 (도형객체로서의 의미를 갖지 않음)	도형 + 위상구조
데이터 형식	DXF	NGI(국토지리정보원 포맷)
기본지리정보구축	별도의 작업공정 필요	기본지리정보 구축의 기반이 되는 데이터

1:1,000과 1:5,000 수치지형도의 제작 및 갱신은 항공측량, 지리조사, 현지보완측량을 통해 디지타이징 방식으로 이루어지지만, 1:25,000 수치지형도는 1:1,000 수치지형도 혹은 1:5,000 지도를 축소 편집하여 제작된다. 이 축소 편집 과정은 별도의 측량이나 지리조사 없이 이미 구축된 대축척 지도를 축소 편집 규정에 따라 편집자가 직접 수작업으로 진행한다. 이 과정에서 수치지형도 축소 편집 관련 규정을 따르게 되는데, 관련 규정으로는 「수치 지도 작성 작업규칙(국토교통부령 제209호)」, 「공공측량 작업규정(국토지리정보원 고시 제2015-2538호)」, 「지도 도식규칙(국토교통부령 제209호)」, 「지형도 도식적용규정(국토지리정보원 고시 제2019-142호)」 등이 있다.

「수치 지도 작성 작업규칙」은 수치 지도 작성의 작업방법 및 기준 등을 정하여 수치 지도의 정확성과 호환성을 확보함을 목적으로 제정되었으며, 공간정보의 표현, 품질검사, 자료 취득 등에 대한 개괄적 내용이 서술되어 있다. 구체적인 내용을 살펴보자면, 제 2조 정의에서 수치 지도 작성 시 사용하는 용어에 대해 정의를 하고 있으며, 제 3조 적용 범위에서는 해당 규칙이 적용되는 범위를 정의하고 있다. 제 4조에서부터 제 8조까지는 수치 지도 작성 시 기준이 되는 좌표계에 관한 내용, 도엽코드 및 도곽의 크기, 수치 지도의 작성 순서, 자료의 취득 등으로 구성되어 있다. 제 9조의 내용은 지형 공간정보의 표현에 관련한 내용으로 축소 편집 시 객체의 표현에 관한 내용과 가장 관련이 높다. 하지만 이 조항에서 규정하고 있는 내용은 지형·지물의 분류체계 및 분류체계의 변동에 관한 내용뿐이다. 제 12조 수치 지도 작성의 세부 기준 항목에서도 수치 지도의 표현에 관련한 내용이 포함되어 있으나, 자세한 내용은 국토지리정보원장이 정한다는 문구로 짧게 서술된 부분 외에 구체적인 지침을 제공하고 있지 않다.

「공공측량 작업규정」은 공공측량 작업계획서 작성기준 등 그밖에 공공측량에 필요한 사항을 정하여 규격을 통일하고 공공측량성과의 정확성을 확보하는 목적으로 제정되었다. 총 7편 188조로 구성되어 앞서 언급한 수치 지도 작성 작업규칙보다 훨씬 많은 양으로 작성되었음을 볼 수 있다. 1편 총칙은 규정의 목적과 정의, 공공측량의 사업 수행 과정 및 관리, 검토에 대한 규칙들에 관한 내용으로 구성되어 있다. 2편은 공공기준점 측량에 관한 내용으로 공공삼각점 측량의 과정 및 관측, 공공수준점 측량의 과정 등에 관한 내용으로 구성되어 있다. 3편 지형측량 부분은 지형도의 작성 과정, 자료 취득 방법 및 그 절차, 실제 지도 제작 과정에 관한 내용으로 구성되어 있다. 특히 제 57조 지도의 축소 편집 부분에서 지도의 축소 편집에 대한 정의와 최종 축소 편집된 지도에 적용되어야 할 규정에 대해 언급하고 있다. 그러나 축소 편집된 지도의 지형지물 표현에 대한 세부지침은 없으며 축소 편집이라는 용어에 대한 정의, 그리고 최종 축소 편집된 지도가 「수치 지도 작성 작업규칙」 또는 「지도도식규칙」 중 관련 규정을 적용해야 한다는 내용만 포함되어 있다. 공공측량 작업규정의 나머지 부분들은 각각 응용측량, 세계 측지계 변환측량, 네트워크 RTK 측량, 기타 응용측량에 관한 내용으로 채워져 있다.

「지도도식규칙」은 간행하는 지도의 도식에 관한 기준을 정하여 지형·지물 및 지명 등을 나타내는 기호나 문자 등의 표시방법 통일을 기함으로써 지도의 정확하고 쉬운 판독에 이바지함을 목적으로 제정되었다. 여기에서 지도의 표현에 관해 언급된 부분은 제 4조에서부터 9조까지의 내용으로 지물을 표현하는 기호와 선의 종류, 등고선, 도곽, 주기의 표시방법과 그 배치 방법에 대해서 정의하고 있다. 그러나 앞서 언급한 규정들과 마찬가지로 지도 축소 편집 시 객체 표현에 대한 세부적인 내용은 포함되어 있지 않다.

지도 축소 편집 시의 지형지물의 표현에 관한 세부 규정은 「지형도 도식적용규정」에 자세하게 다루어지고 있다. 「지형도 도식적용규정」은 지형도 제작에 사용되는 용어를 정의하고 지형도상에 표시되는 기호 및 주기의 취사 선택과 지형지물의 표시방법 및 각종 기호의 적용방법에 관한 기준을 정하기 위한 목적으로 제정되었다. 대상은 1:5,000, 1:10,000, 1:25,000 및 1:50,000 지형도를 대상으로 하며, 이와 관련된 모든 지도 구성요소를 포괄하고 있다. 총 4장 221조의 방대한 내용으로 구성되어 축척별로 어떠한 객체들을 어떻게 표현해야 하는지 규정하고 있다. 제1장에서는 규정의 목적과 대상을 밝히고 규정 내에서 사용되는 용어를 정의하는 한편 다른 규정과의 관계 등을 설명하고 있다. 제2장에서는 다양한 지도의 구성요소들에 대해 각각의 정의, 그리고 표현법에 관해 서술하고 있다. 제3장에서는 주기(annotation)의 표현에 관한 내용으로 구성되어 있다. 주기의 선택, 글자체의 모양과 크기, 축척별 표현 방법 등에 대해 상세히 언급되어 있다. 제4장은 난외사항에 관한 내용으로 구성되어 있다. 지도의 완성도와 독도의 효율성을 높여 사용자가 쉽게 지도를 이해할 수 있도록 도움을 주는 주요사항들의 표현 방법에 관해서 서술하고 있다.

「지형도 도식적용규정」 중 축소 편집 시 객체의 표현 방법에 관한 내용은 제3장 내용에 나타나 있다. 도로의 표시에 관련한 내용은 5절 교통 부분에 언급되어 있으며 1:25,000 및 1:50,000 수치지형도를 대상으로 한다. 해당 규정의 내용을 살펴보면 “고속국도, 일반국도, 국가지원지방도, 지방도, 유료도로 또는 소형차로는 전부 표시한다. 소로 중 마을과 마을을 연결하는 도로, 자동차도와 자동차도 간을 연결하는 도로, 관광목표, 공장, 또는 광산 등에 도달하는 도로, 산림, 습지 등을 통과하는 주요도로 등은 전부 표시하고 기타의 것은 독도의 편의를 고려하여 생략할 수 있다. 시가지 지역, 도로 밀도가 높은 지역 등에 있는 도로는 독도에

필요하다고 인정되는 주요 도로만 표시하고 기타 독도 편의를 고려하여 생략할 수 있다.”라고 언급되어 있다. 이 내용을 간략하게 표로 정리하면 <표 1-3>과 같다.

<표 1-3> 축소 편집 시 도로의 표시에 관련한 규정

도로 유형	삭제 여부
고속국도	X
일반국도	X
국가지원지방도	X
지방도	X
유료도로	X
소형차로	X
소로	O(독도의 편의에 따라)
시가지 지역	O(독도의 편의에 따라)

건물의 표시에 관련한 규정을 살펴보면 “건물의 취사 선택은 밀집 건물 구역에서는 전체의 형태에 큰 변화가 없는 정도까지 생략할 수 있으나 용도상 필요한 독립건물은 생략할 수 없다. 1. 특별시청, 광역시청, 도청 2. 시청, 구청, 군청, 읍, 면, 주민 센터 3. 경찰서, 지구대 또는 파출소, 소방서 등은 독도에 지장이 없는 범위 내에서 우선으로 표시해야 한다.”라고 명시되어 있다. 건물의 표시에 관한 내용은 도로와 비교하면 세부적으로 작성되어 있는데 그 내용을 정리하면 <표 1-4>와 같다.

<표 1-4> 축소 편집 시 건물의 표시에 관련한 규정

건물 유형	삭제 여부
독립건물	돌출부 또는 부속된 작은 창고 등은 생략
밀집 건물	지도상의 짧은 변의 길이가 1.0㎜ 이상인 것을 밀집하여 표시
담	지도상 높이 0.08㎜, 지도상의 길이 3.0㎜ 이상의 가시성이 높은 것만 표시
학교	생략 가능(시가지가 복잡한 경우)
경찰서와 지구대	생략 가능(시가지가 복잡한 경우)
우체국 및 전화국	생략 가능(시가지가 복잡한 경우)
공장	공장 부지면적이 지도상 9.0㎟(정사각형의 경우 한 변의 길이가 3㎜) 미만이면 생략
병원, 보건소	용도상 필요한 것은 표시하고 종합병원과 요양소 등 가시성 또는 인지도가 높은 것은 명칭을 표기. 다만, 보건소는 모두 표기.
관공서	시가지 내에서 표시하기 어려운 것은 일부를 생략

<표 1-3>과 <표 1-4>에서 살펴본 바와 같이 건물과 도로의 축소 편집 시 선택 기준에 대한 확실한 수치가 존재하는 예도 있지만, 도로의 경우 독도의 편의에 따라 삭제 여부를 결정할 수 있고, 건물은 일부 객체에 대하여 세부적인 삭제 기준이 존재하나 많은 경우 기준이 모호하여 상당 부분 편집자의 주관이 반영될 수밖에 없는 현실이다. 편집자의 주관이 반영될수록 지도 일반화의 결과물의 정확성이나 일관성은 담보하기 어려워진다.

기존의 연구결과 보고서(국토지리정보원, 2003, 2014)에서도 이와 같은

축소 편집 제작 과정의 개선이 필요하다고 역설하고 있으며, 현재까지도 대축척 지도로부터 소축척 지도를 자동생성하려고 하는 연구 과제들이 수행되고 있다. 즉, 축소 편집 과정의 개선은 현재의 수치지형도 제작 과정에 있어 매우 중요한 부분 중의 하나라고 할 수 있다.

현재까지의 연구들을 지도 일반화 관점, 기계학습의 관점, 그리고 현재 지도 제작 과정의 관점에서 분석해 본 결과, 지도 일반화 연구에서는 국가지도제작기관을 중심으로 더욱 넓은 지역에 대해 최대한 사람의 개입을 배제하고 자동으로 일반화를 시도하려는 흐름인 것을 볼 수 있었다. 그러나 개선에 앞서 현재 지도 일반화 기술을 적용한 축소 편집된 지도의 결과물이 지도 제작자에 따라 어떤 편차를 보이는지에 대해 분석한 연구는 거의 없는 실정이다. 또한, 현존하는 규정의 미비한 점들에 대한 수정, 보완을 시도하는 연구들은 찾아보기 어렵다.

기계학습 기법의 적용은 지도 제작자 간의 편차를 보다 정확하게 정량화할 수 있다. 최근 연구의 흐름에서도 도로나 집계구 객체의 분류 등에 기계학습을 적용하려는 시도들이 있는데, 본 연구에서도 이 흐름에 맞춰 건물과 도로객체의 선택적 삭제를 대상으로 기계학습 알고리듬을 적용하여 지도 제작자 간의 편차를 정량화하고자 한다. 특히 기계학습 기법을 통해 학습된 모델로 각 제작자가 편집한 지도에 대해 객관적인 정확도 평가가 가능하므로 각 제작자의 편집 결과물에 대한 보다 정확한 분석이 이루어질 수 있다. 또한, 의사결정 나무와 같은 해석 가능한 기계학습 알고리듬의 경우, 알고리듬의 결과를 분석하여 어떤 속성이 선택적 삭제에 주된 영향을 주었는지, 선택의 기준은 어떤지 등에 관한 결과를 도출해 낼 수 있다. 이를 통해 기존 규정의 미비점들을 보완할 수 있을 것으로 기대된다.

1.2.4. 소결

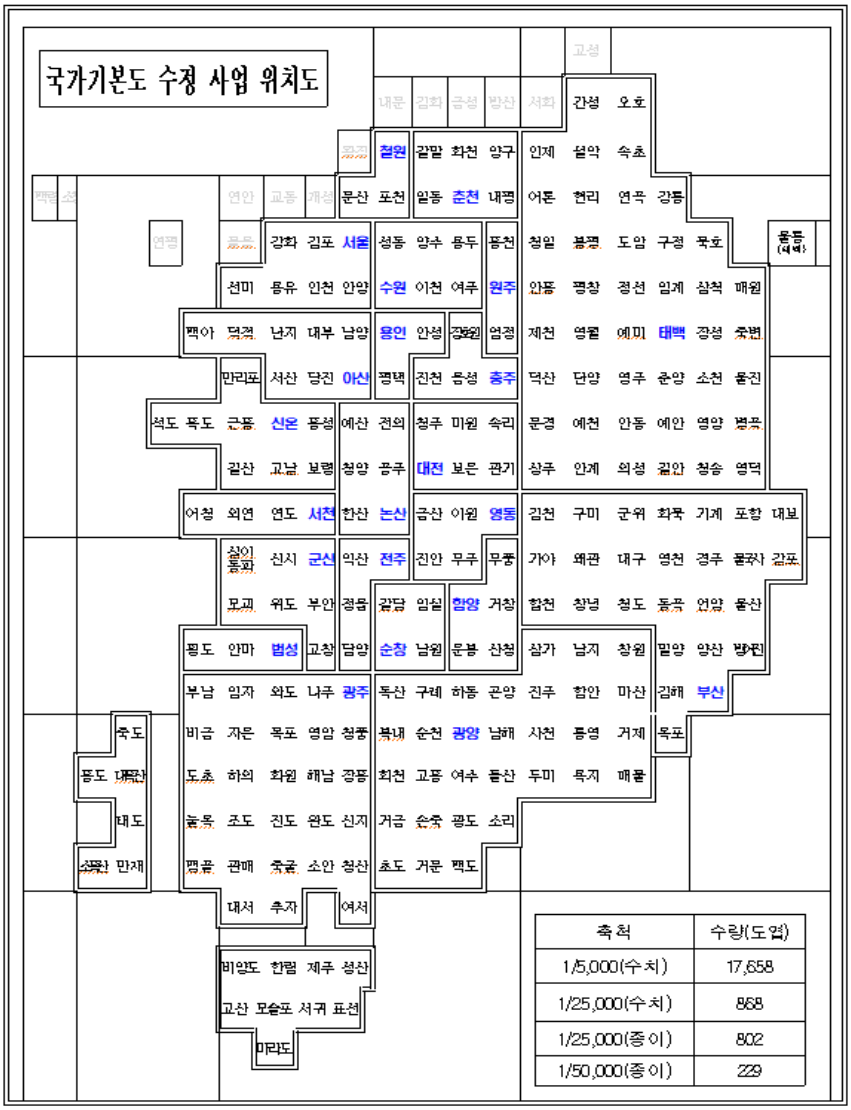
본 장에서는 연구적 측면에서 건물과 도로객체의 지도 일반화에 관한 연구 동향과 기계학습을 활용한 연구 동향을 살펴보고, 실무적 측면에서 국내 지도 제작 및 일반화 과정을 살펴보았다. 지도 일반화에 관한 연구들은 공통으로 사람의 개입 없는 완전한 자동화를 지향하고 있다. 이것은 현재 지도 일반화의 과정에서는 불필요한 사람의 개입으로 인해 과정의 복잡도가 상승하는 한편 정확도 감소 등의 문제가 발생하고 있음을 의미한다. 또한, 실무적인 관점에서도 현재 우리나라의 지도 축소 편집 과정에서 불필요한 공수가 많이 있으며, 뚜렷한 지침이 없어 지도 제작자의 주관이 많이 반영될 수밖에 없는 구조임을 볼 수 있었다. 그러나 지도 일반화 과정에서 사람의 개입으로 발생한 오차들이 어느 정도이며 어떤 양상으로 나타나는지에 대한 연구는 이루어지지 않고 있다. 또한, 기존의 규정 및 지침에서의 미비점들을 수정, 보완하려는 시도 또한 미진한 편이다. 본 연구에서는 기계학습 기법을 활용하여 지도 일반화 과정에서 사람의 개입으로 발생하는 오차의 정도와 그 양상을 정량적으로 규명하고자 한다. 나아가, 도시와 비 도시이라는 두 가지 경우에 대해서 각각 기계학습 모델을 생성하여 최적의 모델을 도출하고 그 모델을 분석하여 미비한 규정 보완에 기여하고자 한다.

1.3. 연구의 범위 및 방법

본 연구에서는 기계학습 기법이 지도 일반화 과정에서 사용될 수 있는 활용 방안을 보이고자 하였다. 이를 위해 수치지형도의 건물과 도로 객체를 대상으로 축소 편집된 결과물에 있어서 지도 제작자 간의 차이를 규명하였다. 또한, 도시지역과 비도시 지역 각각에 대해서 각 지역에 맞는 기계학습 모델을 생성하고 분석함으로써 객체의 선택적 삭제에 영향을 미치는 속성을 분석하고, 나아가 이를 통해 현재의 축소 편집 규정 보완에 기여하고자 한다.

실험 데이터로는 1:5,000 연속수치지형도와 1:25,000 수치지형도의 건물 레이어인 B001 레이어, 도로중심선 레이어인 A002 레이어를 사용하였다. 실험 데이터는 파일 형태 변환 및 좌표체계 매칭, 객체 추출 등의 데이터 가공 과정을 거쳐 실험에 사용하였다. 1:5,000 축척에서 1:25,000 축척으로의 축소 편집은 객체의 선택적 삭제만 적용되고 있으므로 본 연구의 범위 또한 건물과 도로의 선택적 삭제에 대해 기계학습 기법을 적용하는 것으로 하였다.

기계학습 알고리듬의 적용을 위해서 먼저 학습 모델을 생성할 필요가 있다. 학습 모델을 생성하기 위해서 2017년 당시 국가기본도 수정 사업 위치도(<그림 1-5>)의 서로 다른 6개 구획의 지도로부터 데이터를 추출하였다. 현재 우리나라의 수치지형도 축소 편집은 각 구획 별로 서로 다른 제작사에서 담당하고 있으므로 구획이 다른 경우 편집자가 달라지기 때문이다. 건물의 경우 총 180,000개의 데이터, 도로의 경우 120,000개의 데이터를 추출하여 학습 데이터로 활용하였다. 건물 데이터의 경우 각 제작자의 편집 결과로부터 30,000개씩의 건물을 임의 선택하였고, 도로의 경우는 각 20,000개씩 선택하였다.

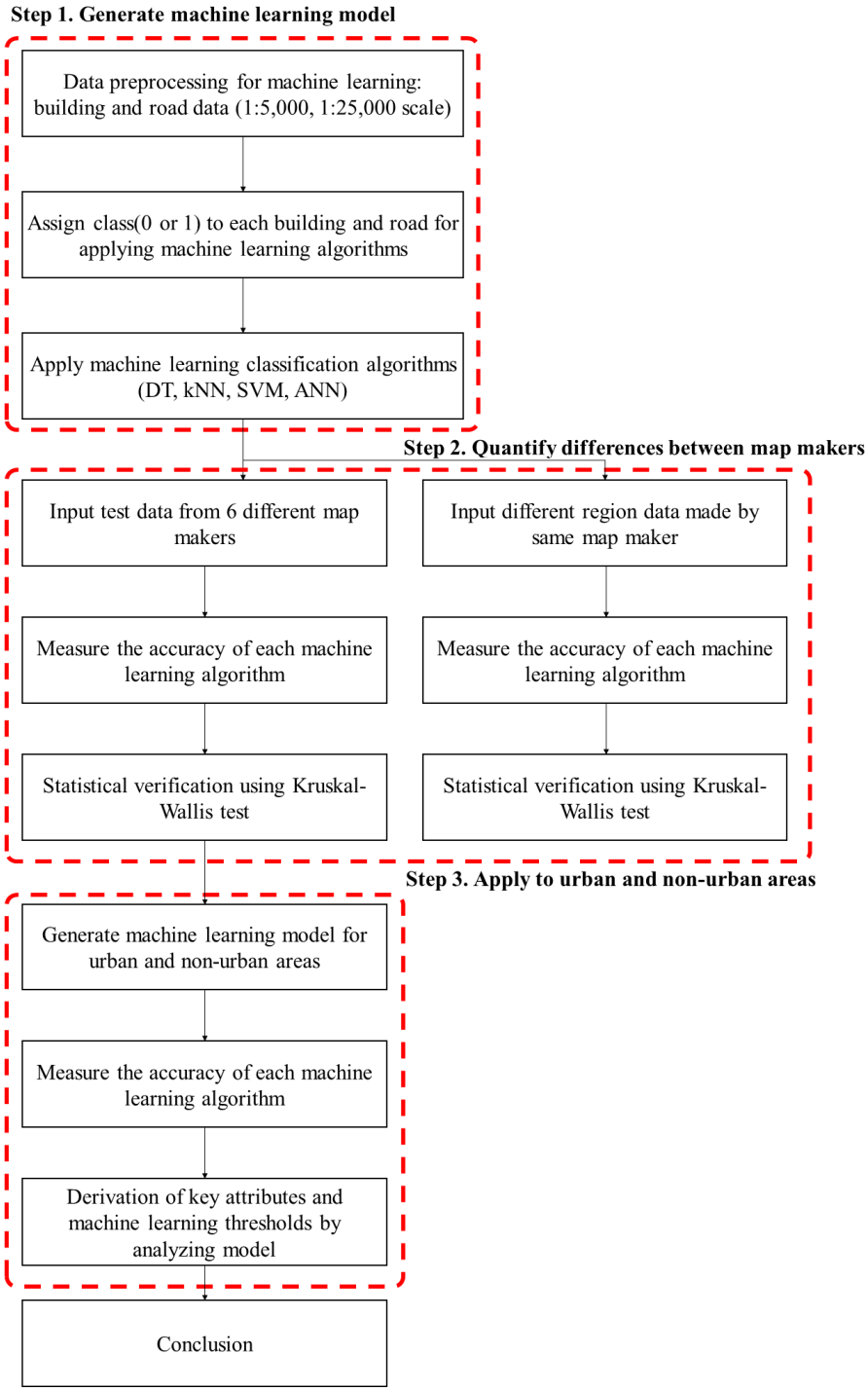


<그림 1-5> 국가기본도 수정 사업 위치도(국토지리정보원, 2017)

지역의 차이가 실험결과에 주는 영향을 최소화하기 위해 데이터 추출 대상 지역을 도시와 비 도시 지역이 혼재되어있는 지리학적으로 유사한 성격을 가지는 지역으로 선정하였다. 또한, 지역별로 균등한 양의 데이터를 임의 추출하여 지역적 변수를 최대한 배제하고자 하였다. 추출한 건

물과 도로 데이터에 대해 각각 기계학습에 필요한 입력 속성과 출력 속성을 정의한 후, 4가지의 기계학습 분류 알고리듬 - 의사결정 나무, k-최근접 이웃, SVM, 인공신경망 - 별로 학습 모델을 생성하였다. 각각의 기계학습 알고리듬별로 학습 모델을 생성한 후에는 서로 다른 6개 제작자가 편집한 6개 지역에 대해 학습 모델을 적용하여 건물과 도로객체의 선택적 삭제에 대한 예측률을 평가하였다. 예측률은 검증 데이터(test data)인 1:5,000 수치지형도의 건물과 도로가 1:25,000 축척에서 삭제되는지를 기계학습 알고리듬을 통해 예측하고, 기계학습 알고리듬의 예측 결과와 실제 작성된 지도와 비교하여 그 값을 측정하였다.

지리적으로 유사한 지역을 선택하고 균등한 양의 데이터를 임의 추출하는 방법으로 지역적인 변수를 최소화하여도 예측률의 편차에는 지역의 차이 때문에 생기는 편차가 혼재되어있다는 문제가 발생한다. 따라서, 같은 제작자가 제작한 서로 다른 지역에 대한 예측률 측정을 통해 지역 간의 차이가 드러난 제작자 간의 편차에 얼마나 영향을 주고 있는지 분석하였다. 또한, 도시지역과 비도시 각 지역에 맞는 기계학습 모델을 생성하고 모델의 분석을 통해 객체의 선택적 삭제에 주된 영향을 주는 속성들을 분석하고, 각 지역 특징에 맞는 축소 편집 규정 개정안을 도출하였다. 위의 과정을 흐름도로 나타낸 것은 <그림 1-6>과 같다. 가장 먼저 학습 모델 생성을 위한 데이터 전처리 과정과 기계학습 알고리듬을 적용하여 모델을 생성하는 과정을 포함하는 기계학습 모델 생성 단계이다. 다음으로는 지도 제작자 간의 편차를 정량화하는 과정이다. 이를 위해 기계학습 기법을 통해 먼저 6명의 서로 다른 지도 제작자의 편차를 밝히고, 해당 편차가 지역의 차이에서 온 것인지를 규명하기 위한 실험을 추가로 진행한다. 마지막으로는 도시와 비 도시 지역에 각각 기계학습 기법을 적용하고, 지역의 특성에 맞는 모델을 생성하는 과정을 진행한다.



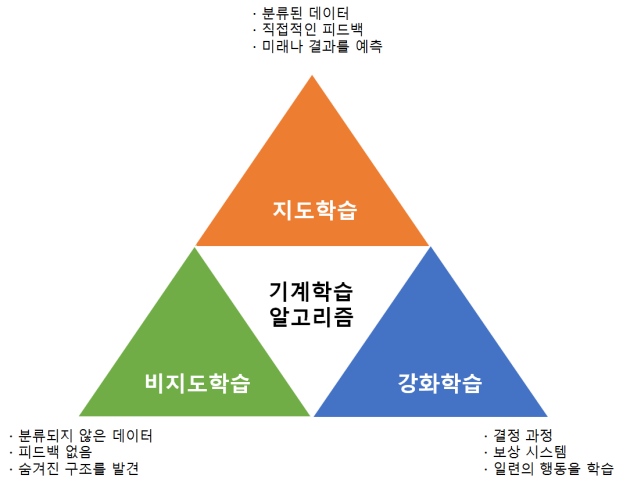
<그림 1-6> 논문 구성 및 흐름도

남은 논문의 구성 순서는 다음과 같다. 2장에서는 기계학습 알고리듬을 소개하고 기계학습 적용을 위한 과정에 관해 서술한다. 3장에서는 실험 및 결과에 관해 서술한다. 먼저 학습 모델 생성 및 생성된 모델을 평가한 결과에 관해 서술하고, 모델 적용을 통한 제작자 간 차이의 정량화를 위한 실험 및 결과에 관해 서술한다. 제작자 간 차이에 대한 실험은 크게 건물과 도로에 대한 평가, 그리고 지역 간의 차이를 평가한 결과에 관해 서술한다. 또한, 도시과 비 도시지역에 대해 각각 기계학습 기법을 적용하고 그 결과 분석을 통한 관련 규정 보완 내용에 관하여 서술한다. 마지막으로 4장에서는 결론 및 고찰로써 논문을 갈음한다.

2. 기계학습 알고리듬의 적용

2.1. 기계학습 개요

본 연구에서는 일반화 과정의 자동화를 위해 지도 일반화에 기계학습 기법을 적용하였다. 기계학습은 본래 컴퓨터 과학의 한 분야로 “명시적으로 프로그래밍 되지 않은 것을 학습할 수 있는 컴퓨터”라고 Samuel(2000)에 의해 정의된 바 있다. 기계학습에서 컴퓨터는 기계학습을 통해 많은 양의 데이터로부터 일반화된 규칙을 도출해 낸다. 기계학습 기술의 핵심은 도출된 규칙을 기반으로 인간에 의해 만들어진 것과 유사한 추론을 만들어 내는 것이다. 이러한 기계학습 알고리듬은 크게 지도학습(supervised learning), 비지도 학습(unsupervised learning), 강화 학습(reinforcement learning)의 세 가지로 나눌 수 있다(<그림 2-1>).



<그림 2-1> 기계학습 알고리듬의 종류
(http://www.techjini.com/blog/machine-learning/)

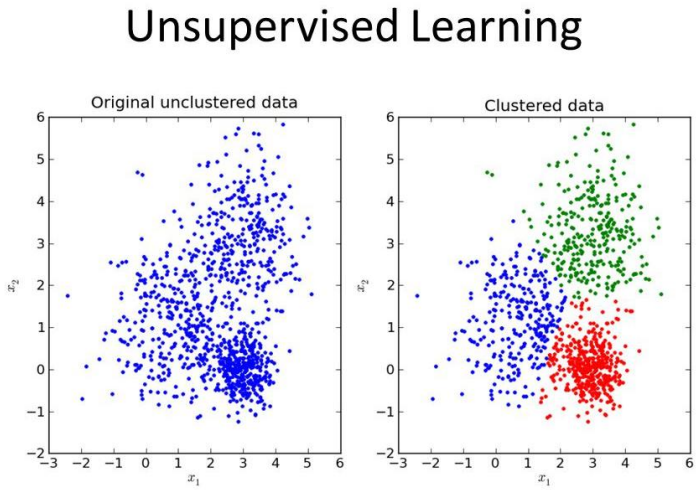
먼저 지도 학습은 데이터에 대한 레이블(label) -명시적인 정답- 이 주어진 상태에서 컴퓨터를 학습시키는 방법이다. 즉, 지도 학습을 위한 훈련 데이터가 입력 객체에 대한 속성을 벡터 형태로 포함하고 있으며 각각의 벡터에 대해 원하는 결과가 무엇인지 레이블로 표시되어 있다. 예를 들어서, 기계학습 분야의 예제로 널리 쓰이는 28×28 크기의 이미지인 MNIST 데이터(<그림 2-2>)의 경우, 훈련 데이터는 다음과 같이 구성된다.

<그림 2-2> MNIST 데이터 세트
(LeCun *et al.,* 2010)

<그림 2-2>와 같이 구성된 훈련 데이터 세트로 학습을 시킨 후, 레이블이 지정되지 않은 테스트 데이터 세트(test data set)를 이용해서 학습된 알고리듬이 얼마나 정확하게 예측하는지를 측정할 수 있다. <그림 2-2>의 MNIST 데이터 세트의 경우, '4'를 나타내는 28×28 이미지)를 학습된 분류기에 입력하였을 때, 올바르게 '4'를 예측하는지, 아니면 '3'이나

‘5’와 같은 잘못된 결과를 예측하는지 측정하는 것이다. 지도 학습 방법으로 예측하는 문제는 그 결괏값의 성질에 따라 크게 분류(classification) 문제와 회귀(regression) 문제로 나눌 수 있다. 분류문제는 예측하려는 결괏값이 이산 값(discrete value)일 때를 의미한다. 예를 들어 MNIST 데이터 세트와 같이 어떠한 이미지에 해당하는 숫자를 예측한다거나 하는 식이다. 또 다른 예로 사진에서 얼굴 인식을 하거나, 음성 인식을 하거나, 또는 이미지 분류를 하는 일 등을 들 수 있다. 한편, 회귀 문제는 예측하려는 결괏값이 연속 값(continuous value)일 때를 의미한다. 대표적인 예로 시간의 흐름에 따른 부동산 가격의 예측 등을 회귀 문제로 볼 수 있다. 지도 학습 방법의 대표적인 알고리듬들에는 분류 문제 알고리듬으로 k-최근접 이웃, 나이브 베이즈(Naïve Bayes), SVM, 의사결정 나무 등이 있고, 회귀 문제 알고리듬으로는 선형 회귀(linear regression), 릿지 회귀(ridge regression), 라쏘 회귀(lasso regression) 등이 있다.

비지도 학습은 지도 학습과는 다르게 데이터에 대한 명시적인 정답을 주지 않은 상태에서 컴퓨터를 학습시키는 방법론이다. 지도 학습 방법이 훈련용 데이터를 통해 데이터를 해석하는 함수를 유추해 낸다고 할 수 있지만, 비지도 학습에서는 그러한 추론이 불가능하다. 따라서 지도 학습에서와 같이 어떠한 결과를 예측하는 데 사용되지 않고 주로 데이터가 어떻게 구성되어 있는지를 밝히는 데 주로 사용된다. <그림 2-3>과 같이 군집화(clustering) 알고리듬이 대표적인 비지도 학습 방법의 알고리듬이라고 할 수 있다.

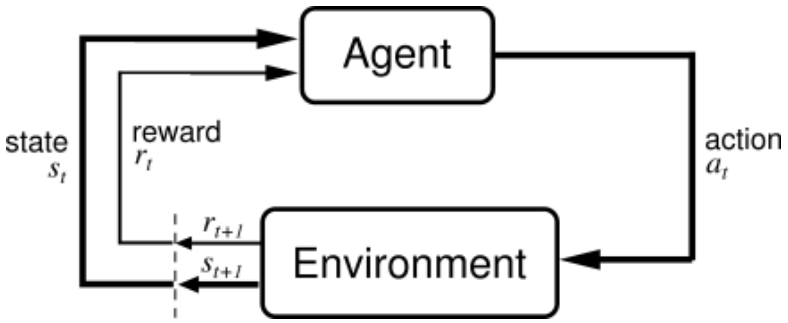


<그림 2-3> 비지도 학습의 예시 - 군집화(clustering)
(알고리듬-처음 배우는 인공지능, 다다 사토시)

비지도 학습의 다른 알고리듬들로는 k 평균(k-means) 알고리듬, 밀도 추정(density estimation) 알고리듬, 기댓값 최대화(expectation maximization) 알고리듬, DBSCAN(density-based spatial clustering of applications with noise) 알고리듬 등이 있다.

강화학습은 지도 학습이나 비지도 학습과는 다른 학습 방법이라고 할 수 있다. 지도 학습과 비지도 학습은 서로 약간의 차이가 있긴 하지만 데이터가 주어진 정적인 상태(static environment)에서 학습을 진행한다는 공통점이 있다. 강화학습은 이와는 다르게 학습의 주체가 되는 에이전트(agent)가 주어진 환경(state)에 대해서 특정한 행동(action)을 취하고, 이로부터 어떠한 보상(reward)을 얻으면서 학습을 진행한다. 이때, 에이전트는 보상을 최대화하도록 학습이 진행된다. 즉, 강화학습은 일종의 동적인 상태(dynamic environment)에서 데이터를 수집하는 과정까지 포함된 알고리듬이라고 할 수 있다. <그림 2-4>는 이러한 강화학습이

진행되는 과정을 보여준다. 에이전트는 환경 s_i에 대한 정보를 받아 행동 a_i를 취하게 되고, 환경으로부터 다음 상태인 s_{i+1}을 확인하고 보상 r_{i+1}을 획득하게 된다.



<그림 2-4> 강화학습 알고리듬
(Sutton and Barto, 2018)

본 연구에서 해결하고자 하는 문제는 대축척 지도의 건물 또는 도로 객체가 소축척 지도에서 남겨지는지 아닌지를 예측하는 문제이다. 이를 위해 기계학습 알고리듬이 지도 제작자의 수동편집 결과를 학습하고 이를 통해 예측을 수행하도록 하였다. 지도 제작자의 편집 결과는 객체가 삭제되었거나 유지되었다는 명시적인 결괏값, 즉 정답이 있는 데이터로 볼 수 있으며, 따라서 지도 객체의 삭제 여부를 예측하는 것은 주어진 정답이 있는 상태에서 기계가 정답에 최대한 가까운 결과를 만들어 내도록 하는 지도 학습의 문제라고 할 수 있다.

2.2. 사용된 기계학습 알고리듬

본 연구에서는 1:5,000 축척에서의 특정 건물이나 도로객체가 1:25,000 축척에서 남겨질 것인지 삭제될 것인지를 기계학습 알고리듬을 통해 추론하고자 하였다. 이를 위해서는 건물들이 삭제 또는 유지되는지를 이진 분류(binary classification)하는 알고리듬의 적용이 필요하다. 대표적인 이진 분류를 위한 여러 알고리듬이 존재하는데, 본 연구에서는 의사결정 나무, k-최근접 이웃, SVM, 인공신경망 총 네 가지 알고리듬을 사용하였다. 각 알고리듬은 이진 분류문제에 적합한 알고리듬들이라고 알려져 있으며, 성능 또한 검증되었다고 알려져 있다(Peto *et al.,* 2008; Svensson, 2016; Naik and Purohit, 2017). 사용된 기계학습 알고리듬들의 특징을 정리하면 <표 2-1>과 같다.

<표 2-1> 사용된 기계학습 알고리듬들의 특징

알고리듬	정확도	훈련속도	결과에 대한 해석	활용성	파라미터 조절 필요성
DT	상대적으로 낮음	빠름	가능	분류/회귀	보통
k-NN	상대적으로 낮음	별도의 훈련 없음	가능	분류/회귀	적음
SVM	상대적으로 높음	빠름 (입력 속성의 개수에 따라 다름)	불확실	분류/회귀	많음
ANN	상대적으로 높음	보통 (입력 속성의 개수에 따라 다름)	불가능	분류/회귀	많음

2.2.1. 의사결정 나무

의사결정 나무 알고리듬은 가장 널리 알려진 기계학습 알고리듬으로 특히 지도 학습에서 유용하게 사용되고 있는 알고리듬이다. 이 알고리듬은 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 의사결정 나무가 학습되는 과정의 핵심은 분류에 있어서 불순도(impurity)를 최소화하는 방향으로 학습이 진행된다. 불순도를 측정하는 대표적인 지수로는 지니계수(Gini index), 엔트로피(entropy) 등이 있으며, 의사결정 나무 알고리듬에 따라 사용하는 지수가 달라진다. 먼저 지니 불순도는 집합에 이질적인 것이 얼마나 섞였는지를 측정하는 지표이며 CART 알고리즘에서 분리 기준으로 사용된다. J클래스를 지니는 요소들의 집합이 있다고 가정할 때, $P(j), j \in 1,2,...J$ 는 집합에서 클래스 j를 지닌 요소를 선택할 확률, 또는 집합 내의 클래스 j를 가지는 요소의 비율로 정의하면 지니 불순도는 수식 (1)과 같이 정의할 수 있다.

$$G = \sum_{j=1}^{J} P_{(j)}(1 - P_{(j)}) \qquad (1)$$

불순도를 측정하는 또 다른 지수인 엔트로피는 확률 변수의 불확실성을 수치로 나타낸 것이다. 엔트로피가 높을수록 불확실성이 높은 것이라고 할 수 있다. m개의 요소가 속하는 A 영역에 대한 엔트로피는 수식 (2)와 같이 정의된다. 여기서 p_k는 A 영역에 속하는 요소 가운데 k 범주에 속하는 요소의 비율을 의미한다.

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 p_k \qquad (2)$$

만약 A 영역에 속한 모든 요소가 같은 범주에 속하면 엔트로피는 0이 되고 반대로 범주가 둘뿐이고 해당 개체의 수가 같게 반반씩 섞여 있으면 엔트로피는 1의 값을 갖게 된다.

의사결정 나무는 나무가 분할된 뒤에 각 노드의 순도(homogeneity)가 증가하고 불확실성이 최대한 감소하도록 하는 방향으로 학습을 진행한다. 하지만 나무의 분할을 과도하게 많이 하다 보면 과적합(over fitting) 문제가 발생하기 쉽게 된다. 이러한 과적합 문제를 방지하는 방법을 가지치기(pruning)라고 한다. 가지치기의 과정은 보통 의사결정 나무를 더는 분할이 이루어지지 않는 상태까지 성장시킨 후, 가장 아래의 가지부터 시작하여 분류 전과 분류 후의 예측 오차를 비교한다. 만약 분류 후의 예측 오차가 더 크다면 가지치기를 진행하게 된다.

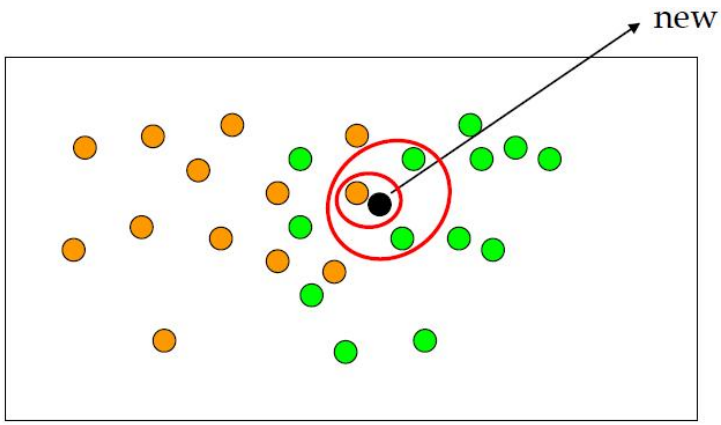
의사결정 나무 알고리듬이 가지는 가장 두드러진 장점은 다른 기계학습 알고리듬과는 달리 결과에 대한 해석이 가능하다는 점이다. 생성된 의사결정 나무의 분할 기준들을 살피면 어떠한 속성들이 분류의 기준이 되고 그 기준값은 얼마인지를 볼 수 있기 때문이다. 본 연구에서도 이러한 장점 때문에 의사결정 나무를 사용하게 되었다. 본 연구의 관련 분야에서 의사결정 나무가 활용된 사례들을 살펴보면 공간적 자기 상관성을 고려한 의사결정 나무를 개발하여 농경지의 생산성을 예측하려는 시도부터 (Li and Claramunt, 2006), 위성영상으로부터 특정 식생을 분류하고 (Liu *et al.,* 2008), 공간분석을 통한 중금속 오염지역을 예측(Liu *et al.,* 2008)하는데 활용하거나, 집계구(면 데이터)의 선택 및 삭제에 대한 예측을 수행(Karsznia and Weibel, 2018) 하는 데 활용되는 것을 볼 수 있었

다. 특히 Karsznia and Weibel(2018)의 연구에서는 다수의 의사결정 나무를 생성하여 각각의 분석을 통해 집계구를 선택하는데 활용될 수 있는 규칙을 도출하려 하였는데 이것은 결과에 대한 해석이 쉬운 의사결정 나무의 특징 때문에 가능한 부분이라고 할 수 있다.

의사결정 나무는 이처럼 이미지 데이터와 벡터 데이터를 가리지 않고 예측 및 분류에 현재까지도 널리 쓰이고 있는 알고리듬이다. 대표적인 의사결정 나무 알고리듬들로는 CART, CHAID, C4.5, ID3 등이 있는데, 이 중 CART 알고리듬은 분할이 항상 이진 분할로 이루어지고, 가지치기가 수월하다는 장점이 있다. 이러한 장점을 활용하기 위해 본 연구에서는 CART 알고리듬을 사용하였다.

2.2.2. k-최근접 이웃

k-최근접 이웃 알고리듬은 고전적인 기계학습 알고리듬으로 새로운 데이터가 입력됐을 때 기존 데이터 가운데 가장 가까운 k개 이웃의 정보로 새로운 데이터를 예측하는 알고리듬이다. <그림 2-5>를 예로 들면, 새로운 개체인 검은 점의 범주를 주변 객체에 따라서 추론하게 되는데, 이때 k=1이라면 주황색, k=3이라면 녹색으로 분류하게 되는 것이다.



<그림 2-5> k-최근접 이웃 알고리듬의 예시

k-최근접 이웃 알고리듬이 다른 알고리듬에 비해 가지는 가장 큰 특징은 별도의 학습 절차가 없다는 것이다. 이것은 k-최근접 이웃 알고리듬이 새로운 데이터가 입력된 이후에야 기존의 데이터들과의 거리를 측정하여 이웃을 추출하기 때문이다. 이러한 의미에서 게으른 모델(lazy model)이라고도 하며, 데이터로부터 모델을 생성해 추론하는 모델 기반 학습(model-based learning)과 대비되는 개념으로 객체 기반 학습(instance-based learning)이라고도 한다.

k-최근접 이웃 알고리듬이 새로운 데이터를 분류하기 위해서는 탐색할 이웃의 개수, 즉 k 값과 이웃과의 거리를 측정하는 방법을 결정해야 한다. 특히 알고리듬의 성능을 위해서 적절한 k 값을 설정하는 것은 대단히 중요하다. k 값이 작으면 데이터의 지역적인 특성이 지나치게 반영되어 과적합 문제가 나타날 수 있다. 반대로 k 값이 커지면 커질수록 분류에 있어서 잡음(noise)의 영향이 줄어들지만, 항목 간 경계가 불분명해지는 문제(under fitting)가 발생할 수 있다. 최적의 k 값은 데이터마다 다르므로 보통 탐욕적(greedy) 방식 - 1부터 꾸준히 증가시키면서 오차

율을 점검하는 방식 - 으로 최적의 k 값을 결정하게 된다.

이웃과의 거리를 측정하는 방법 또한 알고리듬의 결과에 큰 영향을 준다. k-최근접 이웃 알고리듬에서 이웃과의 거리를 측정하는 방법에는 몇 가지가 있다. 대표적으로 유클리드 거리, 맨해튼 거리, 마할라노비스 거리 등이 있는데, 본 연구에서는 변수 간의 상관관계를 반영할 수 있는 마할라노비스 거리 방법을 선택하였다. 건물 데이터의 경우 건물의 면적과 높이, 도로 데이터의 경우 도로 폭과 길이는 변수 간 상관관계가 매우 높고, 분류 결과에 이러한 점들이 반영될 필요가 있기 때문이다. 두 점 사이의 마할라노비스 거리를 측정하는 방법은 수식 (3)과 같다.

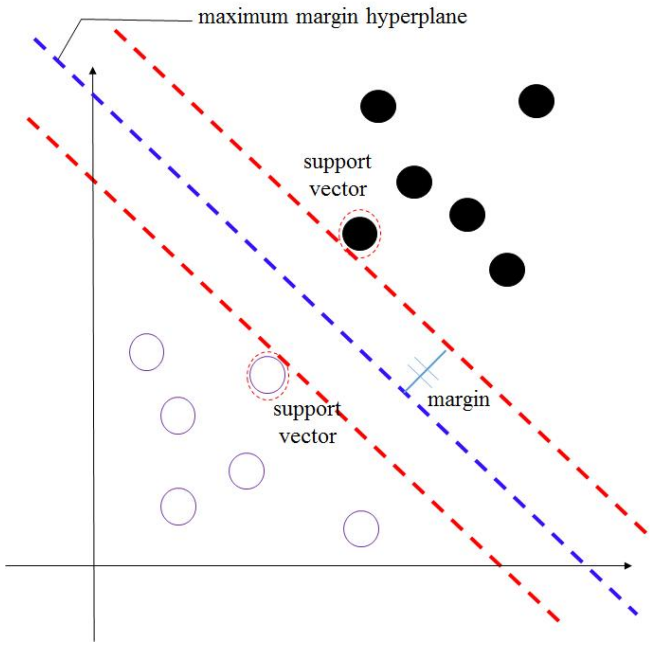
$$d_{Mahalanobis(X,Y)} = \sqrt{(A-B)\Sigma^{-1}(A-B)^T} \qquad (3)$$

k-최근접 이웃 알고리듬은 원리는 단순하지만, 학습 데이터 내에 끼어있는 잡음의 영향을 크게 받지 않으며, 알고리듬의 결과 또한 매우 일관성 있는 결과를 나타낸다는 장점이 있다. 별도의 학습 절차가 없고 새로운 데이터가 입력될 때마다 모든 데이터와의 거리를 매번 측정한다는 점이 이 알고리듬의 가장 큰 특징이자 단점이라고 할 수 있다.

k-최근접 이웃 알고리듬은 현재까지도 널리 사용되는 알고리듬이다. 본 연구의 관련 분야에서는 이동 객체의 궤적을 예측하거나(Güting *et al.*, 2010) 강우로 인한 산사태의 예측(Bui *et al.,* 2017)에 활용되었으며, 토지 피복 분류에 있어 다른 기계학습 알고리듬과 함께 사용하여 정확도를 비교하기 위한 용도로 활용되고 있는 것을 볼 수 있었다(Noi and Kappas, 2018). 따라서 본 연구에서도 k-최근접 이웃을 다른 기계학습 알고리듬들의 성능을 비교할 수 있는 하나의 기준 알고리듬으로써 활용하였다.

2.2.3. SVM

SVM은 분류문제를 해결하는 지도 학습 모델 중 하나이다. 이 알고리듬은 결정 경계(decision boundary)라는 벡터공간 내에 위치하는 데이터들을 가장 잘 분류할 수 있는 데이터 간의 경계를 정의하여 분류를 수행하고, 예측하고자 하는 데이터가 어느 경계면에 속하는지를 확인함으로써 해당 데이터의 클래스를 예측하는 알고리듬이다. <그림 2-6>은 SVM이 데이터를 분류하는 과정을 간단하게 나타낸 그림이다.



<그림 2-6> SVM의 최대 마진 초평면과 서포트 벡터

하얀 점과 검은 점을 분류한다고 할 때, 가운데 파란 점선이 결정경계가 된다. 파란 점선은 단순한 경계인 것이 아니라 최대 마진 초평면(Maximum Margin Hyperplane, MMH)이기도 하다. 3가지 이상의 입력 속성이 존재한다면 분류 경계는 선이 아니라 평면이 될 것이고, 속성의

개수가 늘어날수록 결정경계 또한 고차원이 된다. 이를 초평면(hyperplane)이라고 한다. SVM은 최적의 결정 경계(초평면)를 찾는 알고리듬이라고 할 수 있다. 두 클래스로 구성된 선형 분리가 가능한 문제의 학습 가능한 결정경계는 무수히 많이 존재하나, SVM은 두 클래스의 분리 경계와 근접한 학습 벡터들과 최대 마진(maximal margin) 거리에 있는 벡터 정보를 이용하여 최대 마진 초평면을 찾는 과정인 것이다. <그림 2-6>을 예로 들면 초평면으로부터 데이터까지의 최단 거리가 마진(margin), 그리고 최단 거리에 있는 데이터를 서포트 벡터(support vector)라고 하게 된다. 이 서포트 벡터들이 결국 결정경계를 결정하게 된다.

SVM에서 최대 마진 초평면 $d(x)$를 구하는 방식은 수식 (4)와 같다.

$$d(x) = \omega^T x + b \qquad (4)$$

이 식에서 ω는 초평면과 수직인 법선 벡터가 되고, b는 원점에서 직선까지의 거리를 결정하는 값이 된다. $d(x)$는 데이터가 존재하는 공간을 수식 (5)와 같이 두 영역으로 나눈다. x_1은 <그림 2-6>에서 검은색 서포트 벡터를 의미하고, x_2는 흰색 서포트 벡터를 의미한다.

$$\begin{aligned} d(x_1) &= \omega^T x_1 + \omega_0 > 0 \\ d(x_2) &= \omega^T x_1 + \omega_0 < 0 \end{aligned} \qquad (5)$$

또한, 임의의 점 x에서 초평면까지의 거리는 수식 (6)과 같이 나타낼 수 있다.

$$h = \frac{|d_x|}{\| \omega \|} \qquad (6)$$

SVM에서 풀고자 하는 문제는 위에서 언급했듯이 마진을 가장 크게 하는 초평면을 찾는 것이다. 마진은 서포트 벡터에 의해 결정되며, 서포트 벡터 x에 대한 마진은 수식 (7)과 같이 나타낼 수 있다.

$$margin = \frac{2|d_x|}{\| \omega \|} = \frac{2}{\| \omega \|} \qquad (7)$$

알고리듬을 학습시키기 위한 훈련 데이터 세트를 $X = (x_1, t_1, ..., x_n, t_n)$이라고 할 때, t_i는 데이터의 클래스를 나타내며, w_1에 속하면 $t_i = 1$이고 w_2에 속하면 $t_i = -1$이 된다. 이러한 조건에서 최대 마진을 갖는 초평면을 찾는 것은 조건부 최적화 문제로 나타낼 수 있고, 수식 (8)에서 등호가 성립하는 데이터가 바로 서포트 벡터이다.

$$\max \frac{2}{\| \omega \|} \rightarrow \min \frac{1}{2} \| w \|^2 = \min \frac{1}{2} w^T \cdot w \qquad (8)$$

정리해 보면 SVM은 마진을 최대화하며 위의 제약조건을 만족시키는 ω를 찾는 문제가 된다. 이를 구하기 위해서는 라그랑지안 승수법(Lagrange multiplier method)과 KKT(Karush-Kuhn-Tucker) 조건을 활용하게 된다.

SVM은 크게 하드 마진(hard margin) 방식과 소프트 마진(soft margin) 방식 두 가지로 동작하게 되는데, 하드 마진 방식은 매우 엄격하게 두 개의 클래스를 분리하는 초평면을 구하는 방법이기 때문에, 모든 입력 데이터가 초평면을 사이에 두고 한 클래스에 속해야 한다. 그렇게 되면 몇 개의 잡음으로 인해 두 그룹을 구별하는 초평면이 잘못 구해지거나 아예 구할 수 없게 되는 문제가 발생하게 된다. 소프트 마진 방

식은 이것을 해결하기 위해 나온 방식으로 기본적으로 하드 마진 방법을 기반으로 하지만, 차이점은 서포트 벡터가 위치한 경계선에 약간의 여유 변수(slack variable)를 두는 것이다. 현실적으로는 하드 마진 방식을 적용하여 분류를 수행하는 것이 어려우므로 거의 소프트 마진 방식을 사용하게 된다.

SVM은 새로운 데이터가 입력되었을 때, 전체 데이터와의 거리 또는 유사도를 계산하는 것이 아니라, 서포트 벡터와의 거리만 계산하면 되기 때문에 계산 비용을 상당히 줄일 수 있다는 강점이 있다. 특히 이진 분류의 문제를 해결하는 데 있어서 우수한 성능을 보인다고 알려져 있다.

이러한 우수성 때문에 현재까지도 기계학습을 적용한 연구들에서는 가장 널리 쓰이고 있는 알고리듬이기도 하다. 대부분 산사태나 홍수 등 자연재해를 예측하는 데 활용되고 있는 모습이 보인다(Xu *et al.,* 2012; Pourghasemi *et al.,* 2013; Tehrany *et al.,* 2015; Xiong *et al.,* 2019). 자연재해들은 현상에 영향을 미치는 변수의 개수가 매우 많고 복잡한 양상을 띠는데, 이러한 데이터의 경우 SVM이 우수한 성능을 나타내기 때문으로 생각된다. 본 연구에서도 최대한 많은 변수를 활용하여 객체의 선택적 삭제를 예측하려고 시도했기 때문에 SVM 알고리듬을 활용하게 되었다.

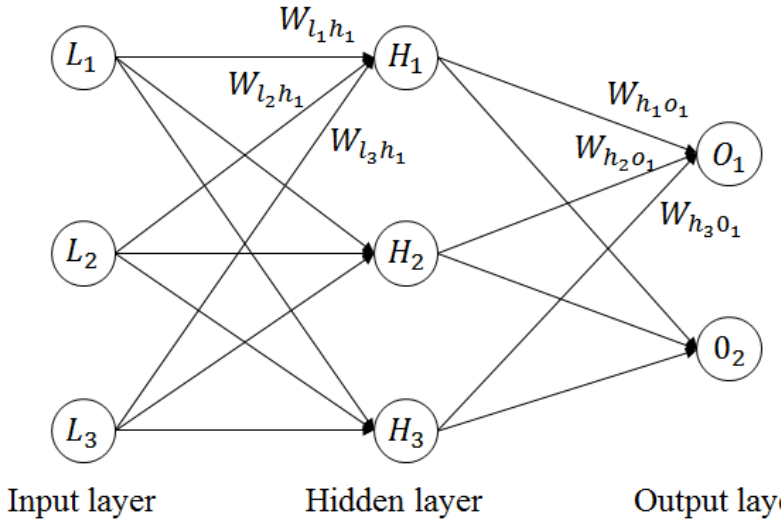
2.2.4. 인공신경망

인공신경망은 사람의 신경망(뉴런)이 학습하는 과정을 묘사한 알고리듬이다. 인공신경망을 구성하는 최소 단위는 퍼셉트론(perceptron)이다. 퍼셉트론은 입력값과 출력값을 가지며, 각 입력에 고유한 가중치(w, weight)가 곱해진다. 가중 합을 구한 후에는 활성 함수를 적용하여 결과

를 출력하게 된다. 입력값 x_i에 대한 출력값은 수식 (9)와 같이 계산된다.

$$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{n} W_i x_i + b) \qquad (9)$$

W는 가중치 w_i의 행렬, b는 편향 값(bias)을 나타낸다. 활성 함수 f는 각 퍼셉트론의 출력 여부를 결정해주는 함수로 대표적으로 Sigmoid, Relu 함수 등이 있다. 이러한 퍼셉트론들을 여러 층 쌓아서 만든 분류기를 다층 퍼셉트론(Multi-Layer Perceptron, MLP) 혹은 신경망(Neural Network, NN)이라고 부른다. 신경망은 <그림 2-7>과 같이 입력층, 은닉층, 출력층으로 구성되어 있다.



<그림 2-7> 뉴런의 값 계산 과정

<그림 2-7>의 신경망의 각 뉴런의 값이 계산되는 과정은 수식 (10), (11)과 같이 나타낼 수 있다.

$$H_1 = L_1 \times w_{l_1 h_1} + L_2 \times w_{l_2 h_1} + L_3 \times w_{l_3 h_1} \quad (10)$$

$$O_1 = H_1 \times w_{h_1 o_1} + H_2 \times w_{h_2 o_1} + H_3 \times w_{h_3 o_1} \quad (11)$$

위 수식과 같이 인공신경망에서 결괏값을 얻기 위해서는 입력층에서부터 출력층까지 차례로 값을 계산하여 전달하게 된다. 이 과정을 전방 전달(feed forward)라고 한다. 신경망이 최적의 결과를 도출하기 위해서는 가중치를 조절하는 과정이 필요하다. 최적의 가중치를 찾는 과정은 비용함수(cost function)를 최소화하는 과정이라고 할 수 있는데, 가장 널리 활용되는 MSE(Mean Square Error)를 활용한 비용함수를 나타내는 방법은 수식 (12)와 같다.

$$\mathrm{cost} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (12)$$

비용함수의 우변의 $\hat{Y}_i$는 신경망을 거쳐서 출력된 예측값이고, Y_i는 실제 클래스의 값이다. 비용함수가 최소화된다는 것은 결국 예측값과 실제 값의 차이가 최소화된다는 것으로 볼 수 있다. 훈련 데이터를 활용하여 가중치와 편향 값을 변화시키는 과정을 반복적으로 수행하여 비용함수가 최솟값이 되도록 하는 것이 신경망 학습의 목표이다. 이 과정에서 주로 경사 하강법(gradient descent method)을 활용하게 된다. 경사 하강법은 비용함수 C를 편미분을 수행하면서 그 미분 값(gradient)이 음이 되는 방향으로 갱신을 반복하다 보면 최적값에 도달하게 되는 방법이다. 경사 하강법을 활용하여 출력 부분에서 입력 부분 방향으로 순차적으로 비용함수에 대한 편미분을 구하고, 얻은 편미분 값을 이용해 가중치와 편향 값을 갱신시킨다. 모든 훈련 데이터에 대해서 이 작업을 반복적으

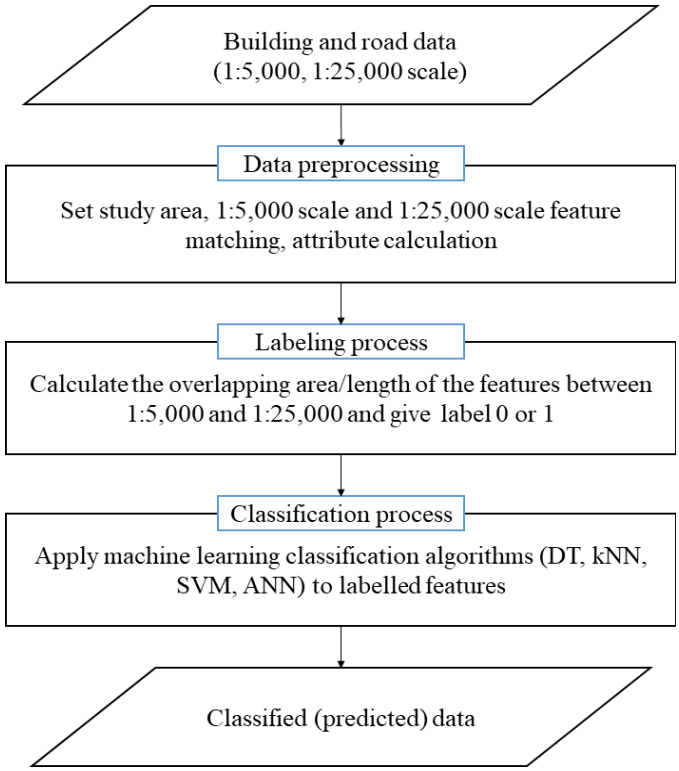
로 수행을 하다 보면, 훈련 데이터에 최적화된 가중치와 편향 값을 얻을 수 있다. 이것을 역전파(back propagation) 알고리듬이라고 한다.

인공신경망 알고리듬은 기존의 기계학습 알고리듬과 비교하면 예측률이 높게 나타나는 것으로 알려져 있다. 그렇지만 결괏값이 어떠한 과정을 거쳐 도출되었는지에 대한 설명하기에는 어려움이 있으며, 이 때문에 종종 블랙박스(black box) 모델로 불리기도 한다. 인공신경망 알고리듬 또한 SVM과 마찬가지로 자연재해를 예측하거나(Biswajeet and Saro, 2007; Pradhan *et al.,* 2010; Choi *et al.,* 2012; Kia *et al.,* 2012) 복잡한 도시 현상을 시뮬레이션(Li and Yeh, 2002; Pijanowski *et al.*, 2014)하는데 활용되고 있다. 또한, 인공신경망 알고리듬은 현재 널리 사용되는 딥러닝 연구의 기본 알고리듬이기도 하다. 본 연구에서는 다양한 변수들을 활용한 예측에 효과적이라는 점과 딥러닝을 활용의 가능성을 검증하려는 측면에서 인공신경망 알고리듬을 사용하였다.

2.3. 기계학습 알고리듬 적용을 위한 데이터 생성

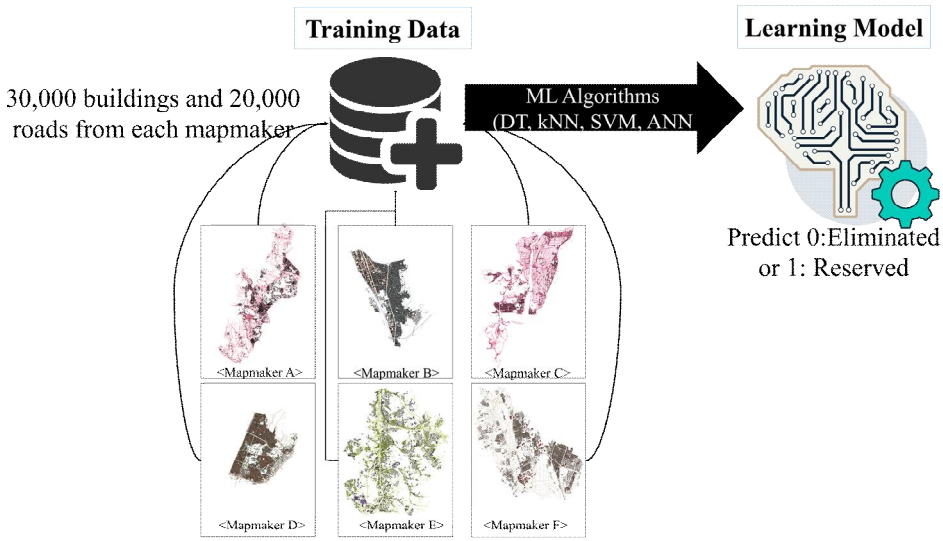
2.3.1. 실험 데이터 및 전체 실험 순서

지도 객체의 선택적 삭제에 기계학습 기법 적용을 위해 데이터의 가공 및 전처리는 필수적이다. 본 연구에서는 먼저 1:5,000 수치지형도와 1:25,000 수치지형도의 건물과 도로의 객체 매칭과 속성 계산을 통해 전처리를 수행하고, 중첩 면적 비 등을 활용하여 클래스를 부여하였다. 이후 4가지의 기계학습 알고리듬을 적용하여 각 모델을 통하여 선택적 삭제 여부를 예측할 수 있도록 하였다. 그 순서는 <그림 2-8>과 같다.



<그림 2-8> 실험 순서도

앞서 1.3 장에서 서술한 것과 같이 기계학습 알고리듬의 모델 생성을 위해 6개의 지도 제작자들이 제작한 6개 지역으로부터 건물 각 30,000개, 도로 각 20,000개씩을 임의 선택하여 건물의 경우 총 180,000개, 도로의 경우 총 120,000개의 데이터를 학습 모델의 훈련 데이터로 사용하였다 (<그림 2-9>).



<그림 2-9> 실험 데이터 생성 과정

1:5,000과 1:25,000 수치지형도는 각각 ESRI Shape file format (.shp)와 Drawing Exchange Format (.dxf)으로 작성되었다. 효율적인 데이터의 처리를 위해, 먼저 두 가지 데이터 포맷을 통일할 필요가 있었고, 따라서 1:25,000 축척 지도를 Shape file 포맷으로 변환하였다. 또한, 1:5,000 수치지형도는 GRS 80/UTM-K 좌표계를 사용하고 있고, 1:25,000 수치지형도는 GRS 80/TM 중부/동부/서부 좌표계를 사용하고 있어서, 먼저 GIS 도구를 활용하여 두 데이터 세트의 좌표계를 일치시킨 후 데이터 처리를 시작하였다. 훈련 데이터는 7:3의 비율로 나누어서 각

각 훈련 데이터와 검증 데이터로 사용되었다. 이와 같이 생성된 데이터를 사용하여 각 기계학습 모델 자체의 성능을 평가하였다. 또한, 각각 6개의 각 지역에서 훈련 데이터로 사용되지 않은 나머지 데이터들은 제작자 간 차이를 규명하기 위한 검증 데이터로써 사용되었다. 즉, A 지역의 경우 건물 55,322개, 도로 30,281개의 객체가 존재하였는데, 훈련 데이터로써 사용된 건물 30,000개와 도로 20,000개를 제외한 건물 25,322개와 도로 10,281개의 데이터가 검증 데이터로써 사용되었다. 또한, 임의 추출된 훈련 데이터를 다시 도시지역과 비도시 지역으로 구분하여 각 지역에 대한 기계학습 모델을 생성하였다. 데이터의 객체별 개수는 <표 2-2>와 같다.

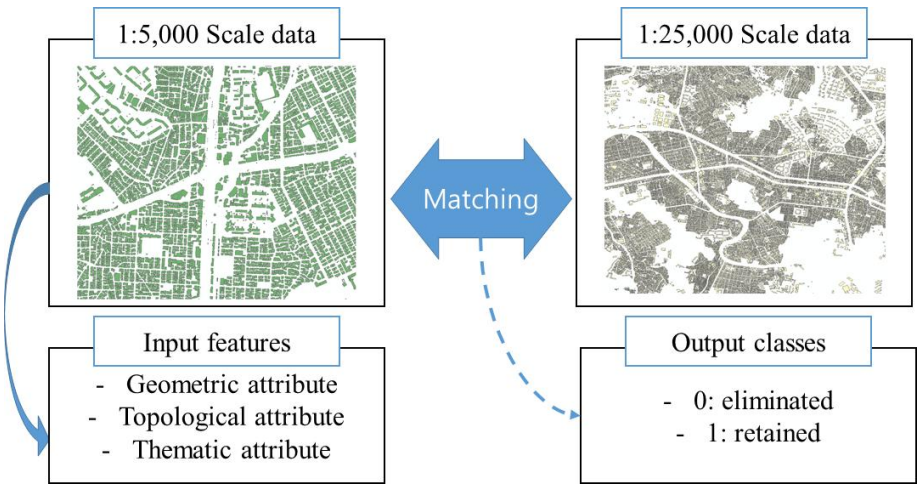
<표 2-2> 제작자별 건물과 도로의 객체 수

지도 제작자	총 객체
제작자 A	건물: 37,611 도로: 22,520
제작자 B	건물: 43,982 도로: 32,416
제작자 C	건물: 52,798 도로: 33,055
제작자 D	건물: 38,232 도로: 24,783
제작자 E	건물: 38,878 도로: 23,640
제작자 F	건물: 42,885 도로: 28,889

2.3.2. 훈련 데이터 생성

기계학습 방법의 적용을 위해서는 먼저 훈련 데이터를 구축할 필요가

있다. 훈련 데이터는 반드시 입력 속성과 출력 클래스를 포함하고 있어야 한다. 본 연구에서는 1:5,000 수치지형도의 속성 데이터로부터 입력 속성을 만들고, 1:5,000과 1:25,000 객체의 매칭을 통해 출력 클래스들을 생성했다(<그림 2-10>).



<그림 2-10> 기계학습을 위한 입/출력 데이터 생성

다음으로는 객체의 특징들로부터 분류를 위한 입력 속성을 정의할 필요가 있다. 입력 속성으로는 우선 건물의 경우 크게 3가지 특성, 즉 기하학적 특성, 위상학적 특성, 의미적 특성을 추출하여 사용하였다. 기하학적 특성으로는 건물의 면적, 둘레, 그리고 높이를 사용하였다. 기하학적 특성은 건물의 선택 여부에 가장 큰 영향을 끼치는 요소로 알려져 있으나(Li *et al.*, 2004) 세 특성 모두 수치지형도 건물 레이어의 속성에 포함되어 있지 않았기 때문에 별도의 계산 과정을 거쳐 속성을 생성하였다. 위상학적 특성은 건물과 다른 지물 사이의 관계에 대한 특성으로 정의하여 사용하였다. 이 특성은 건물의 군집화에 큰 영향을 주는 것으로 알려졌지만, 소규모 건물의 삭제에도 영향을 주는 것으로 알려져 있다

(Damen *et al.,* 2008). 위상학적 특성은 특정 건물에서 가장 인접한 건물까지의 거리와 가장 인접한 교차로와의 거리를 계산하여 사용하였다. 마지막으로 의미적 특성은 건물의 주기, 그리고 주거시설, 상업 시설 등 건물의 용도를 입력 속성으로 사용하였다. 그러나 수치지형도 건물 레이어의 기본 속성은 위와 같은 속성을 전부 포함하고 있지 않기 때문에 별도의 계산 과정을 거쳐서 속성을 추출하였다. 이 과정에서 건물의 이름은 해당 건물의 주기가 있는지 없는지를 판단해 "0 : 주기 없음, 1 : 주기 있음"의 이진(binary) 속성으로 변환하여 사용되었다. 건물 용도는 범주(category) 변수의 값으로 변환하여 사용하였다. 사용된 속성들의 상세한 내용은 <표 2-3>과 같다.

<표 2-3> 사용된 입력 속성(건물)

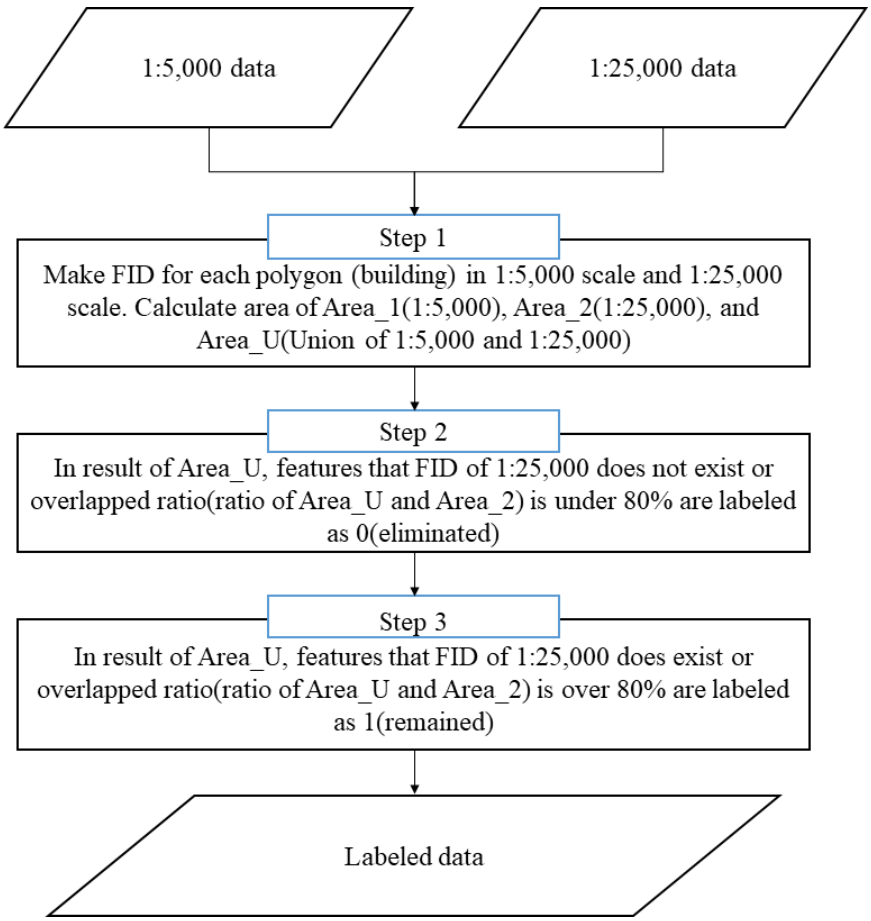
속성명	속성내용	속성타입
건물 명칭	수치지형도의 건물 명칭	String
건물용도	수치지형도의 건물용도	String
건물 면적	각 건물의 면적	Double
건물 둘레	각 건물의 둘레길이	Double
건물의 높이	각 건물의 높이 (건물의 층수 × 3m로 계산)	Double
인접한 건물과의 거리	각 건물로부터 가장 가까운 다른 건물과의 거리	Double
인접한 교차로와의 거리	각 건물로부터 가장 가까운 교차로와의 거리	Double

출력 클래스는 소축척(1:25,000)에서의 건물의 잔존 여부에 따라 "0 : 삭제됨", "1 : 유지됨" 두 가지 클래스가 사용되었다. 1:5,000 수치지형도

와 1:25,000 수치지형도 건물 레이어의 모습은 <그림 2-11>과 같다. 면적이 작은 부속 건물들이 대부분 삭제된 것을 볼 수 있다.

<그림 2-11> 1:5,000 건물 레이어(좌) 와 1:25,000 건물 레이어(우)

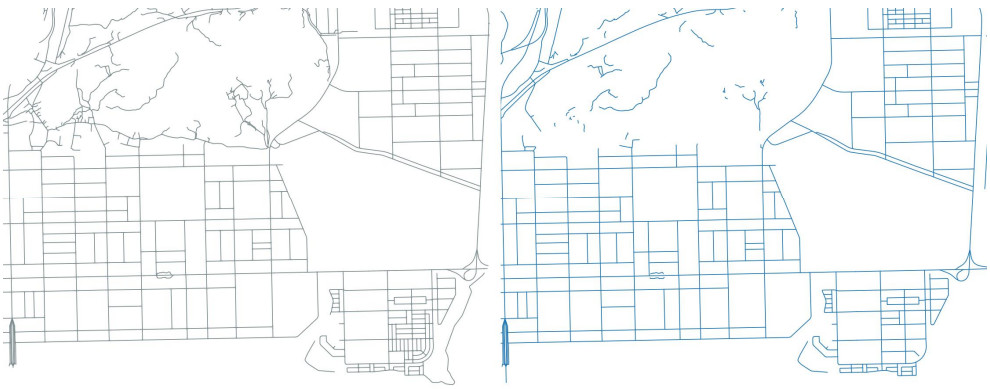
건물에 클래스를 부여하는 과정은 1:5,000 건물과 1:25,000 건물의 중첩 면적 비를 활용한 방법을 사용하였다. 두 데이터 세트에서의 건물이 같은 건물인지를 판단하는 기준은 서로 다른 공간 객체의 유사도 계산에 따라 중첩 면적 비가 80% 이상인 건물로 결정하였다(김지영 등, 2013 ; 박슬아 등, 2014). 클래스 부여 과정은 <그림 2-12>와 같다.



<그림 2-12> 건물의 클래스 부여 과정

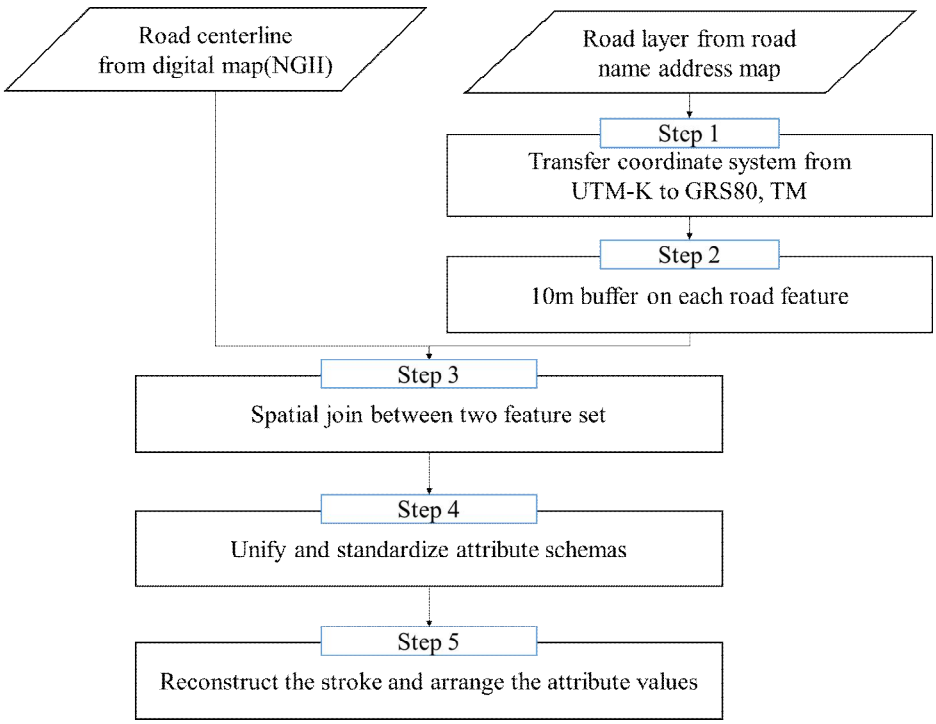
건물의 경우와 마찬가지로 도로 레이어에도 기계학습 방법을 적용하기 위해서는, 먼저 대축척 지도와 소축척 지도의 비교를 통해 대축척에서의 도로객체가 소축척에서 남겨졌는지 아닌지를 판단하는 과정, 즉 클래스를 부여하는 과정이 선행되어야 한다. 문제는 건물과는 다르게 도로는 객체 대 객체 단위로 비교하기가 매우 어렵다는 점이다. 대략적인 1:5,000 수치지형도와 1:25,000 수치지형도 도로 레이어의 모습은 <그림 2-13>과 같다. 소로나 이면도로, 진입로 등이 대부분 삭제된 것을 볼 수

있다. 수치지형도의 도로중심선 데이터는 도로의 모든 교차점을 기준으로 개별적인 도로객체로 나누어져 있으며, 교차점 사이에서도 여러 개의 객체로 쪼개어진 경우가 많다. 문제는 쪼개어져 있는 정도가 1:5,000 지도와 1:25,000 지도 사이에서 차이가 있으므로 도로객체별로 비교하여 클래스를 부여하는 과정을 바로 적용할 수가 없다. 따라서 구별된 클래스 부여를 위해, 서로 다른 두 축척에서의 도로 레이어를 하나의 체계로 통일할 필요가 있다.



<그림 2-13> 1:5,000 도로중심선 레이어(좌) 와 1:25,000 도로중심선 레이어

이를 위해 본 연구에서는 박우진(2013)의 연구에서 사용되었던 도로명주소 전자지도의 도로 구간 레이어를 활용하여 같은 도로명을 갖는 구간을 하나의 단위로 재구성하는 방법을 채택하였다. 수치지형도와는 달리 도로명주소 전자지도의 도로 구간은 도로명 단위로 하나의 객체를 이루고 있으므로 이러한 재구성이 가능하다. 수치지형도 도로중심선의 재구조화를 위한 프로세스는 <그림 2-14>와 같다.

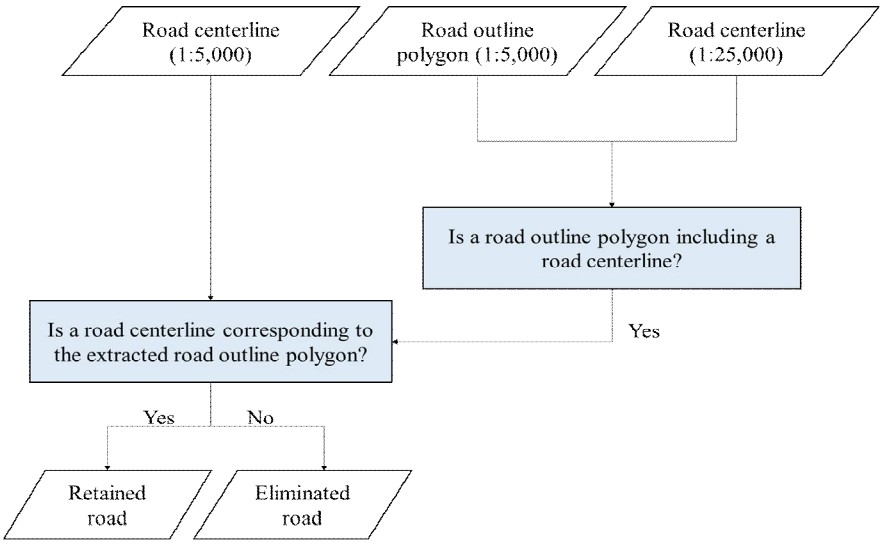


<그림 2-14> 수치지형도 도로중심선 재구조화 과정

먼저, 도로명주소 전자지도의 도로 구간 레이어와 수치지형도의 좌표체계가 상이하므로, 좌표변환을 통해 두 레이어의 좌표체계를 통일시킨다. 그 후, 도로명주소 전자지도의 도로 구간 레이어에 버퍼를 적용하여 수치지형도의 도로중심선이 포함되도록 한다. 이를 위해서는 적절한 버퍼 폭이 적용되어야 하는데, 본 연구에서는 실험적(heuristic)으로 10*m*를 적용하였다. 버퍼 폭을 너무 크게 하면 한 개의 버퍼에 여러 개의 도로중심선이 포함되어 매칭에 어려움이 있을 수도 있고, 버퍼 폭을 너무 작게 하면 도로중심선이 버퍼 구간 내에 포함되지 않을 수도 있는데, 여러 가지 버퍼 폭을 적용해본 결과 10*m* 정도에서 오류 없이 처리되었기 때문이다. 적용된 버퍼 영역과 도로중심선에 대하여 공간조인(spatial join)을 수행하여 도로중심선 객체를 공간적으로 포함하고 있는 버퍼객체를

매칭 하여 그에 해당하는 도로 구간의 속성들(도로명, 도로 위계, 도로 폭 등)을 해당 도로중심선 객체의 속성값에 추가시킨다. 이때 수치지형도의 도로중심선 레이어와 도로명주소의 도로 구간 레이어에 각각 같은 값을 갖는 속성들(도로 폭, 도로 길이 등)의 스키마를 통일시킨 후, 도로 구간을 기준으로 재구성하는 것으로 전처리 과정을 종료한다.

전처리 과정이 완료된 후, 1:5,000에서의 도로가 1:25,000에서 남아있는지 아닌지를 판단하여 “0 : 삭제, 1 : 유지”의 속성을 부여한다. 이때, 판단의 기준은 널리 사용되는 버퍼 폴리곤(buffer polygon)을 활용한 방법을 사용한다(Zhang, 2009). 먼저 1:5,000 도로중심선 레이어로부터 생성된 버퍼 폴리곤과 1:25,000 도로중심선 레이어를 중첩한다. 이 과정에서 1:25,000 도로중심선을 포함하는 1:5,000 도로중심선 버퍼 폴리곤을 추출한다. 그리고 추출된 버퍼 폴리곤에 포함되는 1:5,000 도로중심선은 남겨진 객체로, 포함되지 않는 도로중심선은 제거된 객체로 분류한다. 버퍼 폴리곤을 이용한 판단 과정은 <그림 2-15>와 같다.



<그림 2-15> 버퍼 폴리곤을 활용한 도로중심선 레이어 분류 과정

전처리가 완료된 후 건물과 마찬가지로 입력 속성으로 사용될 특징들을 정의하였다. 도로의 경우는 수치지형도 속성자료로 존재하는 도로 구분, 포장 재질, 분리대 유무, 차로 수, 도로 폭 속성을 사용하였으며, 도로 길이, 도로의 위계 속성을 추가로 부여하여 입력 속성으로 사용하였다. 도로의 속성들은 Zhou and Li(2014)의 연구에서 도로의 선택적 삭제 시 사용한 속성들을 참고하여 선정하였다. 사용된 속성들의 상세한 내용은 <표 2-4>와 같다.

<표 2-4> 사용된 입력 속성(도로)

속성명	속성내용	속성타입
도로 구분	수치지형도 도로 구분	String
포장 재질	포장, 비포장	String
분리대 유·무	분리대 유, 무	String
차로 수	동일 ID 내의 도로객체들의 차로 수 평균값	Long
도로 폭	동일 ID 내의 도로객체들의 도로 폭 평균값	Double
도로 길이	동일 ID 내 도로객체들의 도로 길이 총합	Double
중요도	도로의 연결성 고려	Double

3. 실험 및 결과

3.1. 건물과 도로에 기계학습 적용

3.1.1. 건물에 적용

기계학습의 적용을 위해 7개의 입력 속성과 2개의 출력 클래스를 지정하였다. 입력 속성은 건물의 주기 여부, 건물의 높이, 건물의 면적, 건물의 둘레, 가장 가까운 건물과의 거리, 가장 가까운 교차로와의 거리, 그리고 건물의 종류가 사용되었고, 출력 클래스로는 위에 언급된 바와 같이 “0 : 제거됨, 1 : 유지됨”으로 구분하여 사용하였다. 다른 기계학습 알고리듬의 적용 사례에서는 입력 속성 중에 일부만 선별하여 사용하기도 하는데, 본 연구에서는 입력 속성의 수가 많지 않기 때문에 모든 속성을 다 사용하였다. 건물의 면적이나 둘레와 같은 일부 속성값들은 속성값 간의 편차가 심하므로 정규화가 필요하다. 정규화는 수식 (13)과 같이 진행되었다.

여기에서 μ는 속성들의 평균값이며, σ는 속성들의 표준편차이다.

$$x_j' = \frac{x_j - \mu_j}{\sigma_j} \tag{13}$$

학습 모델 생성을 위해 서로 다른 6명의 지도 제작자가 편집한 지역으로부터 각 30,000개씩의 건물을 추출하여 180,000개의 건물 데이터를 사용하였다. 실험 데이터의 테이블 구조는 <표 3-1>과 같다.

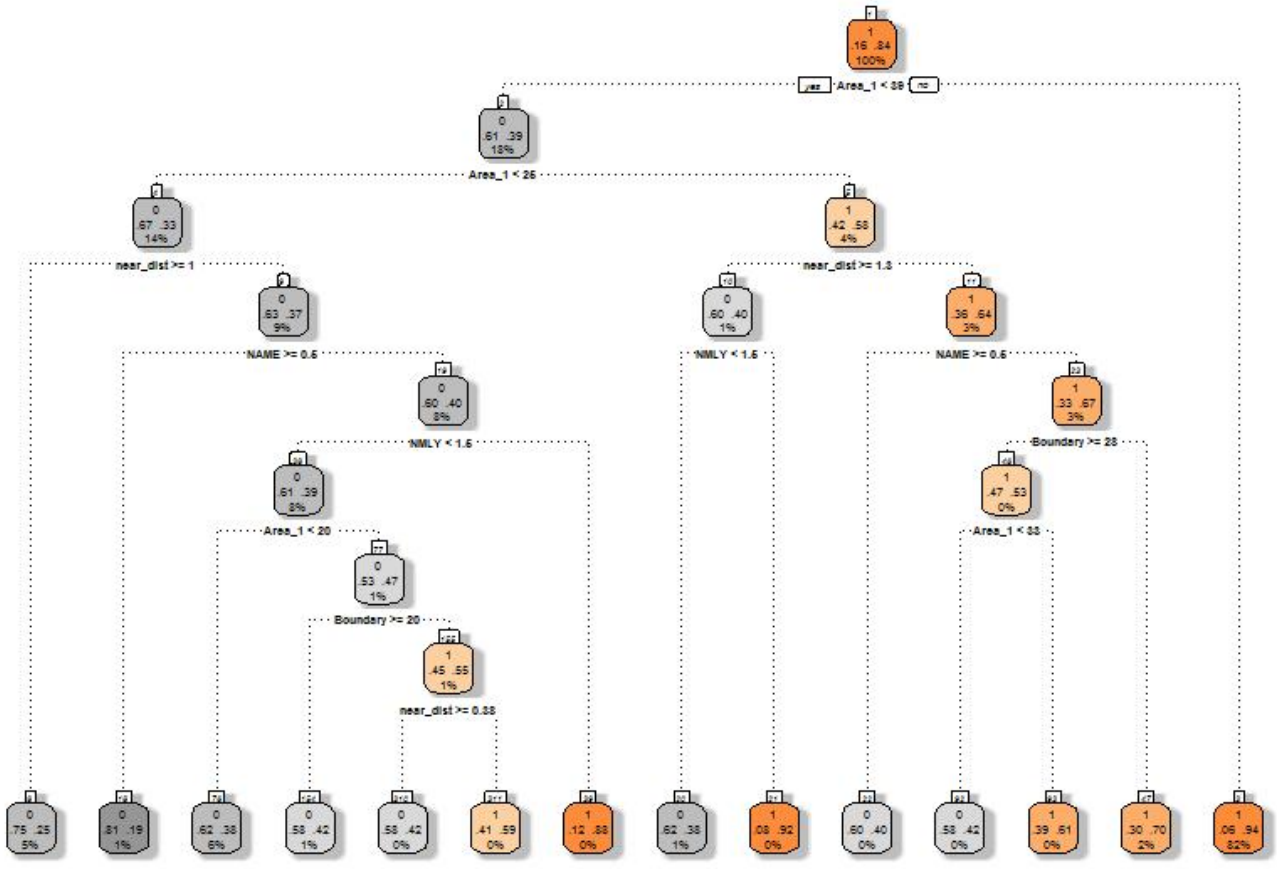
<표 3-1> 건물의 실험 데이터 구조 예시

ID	usage	annotation	floor	area	boundary	near_junction	near_build	index
0	1	1	3	184.7487	57.13371	3.089799205	0.078605944	1
1	2	1	4	165.8014	56.80402	45.39850064	1.497363961	0
2	1	1	2	161.2139	56.74296	0.771216732	0.00420016	1
3	1	0	3	133.7038	46.40596	1.120453705	0.13616514	1
4	3	0	2	86.98367	38.70486	495.937875	0.691803111	1
5	1	1	2	142.0842	57.69059	22.21556896	0.276461345	1
6	1	1	5	152.1676	50.37416	2.255310261	1.769402566	1
7	3	0	2	115.1654	46.48673	33.79661844	0.861486587	1
8	3	0	2	99.04999	43.03149	89.26863248	0.07271393	1
9	3	1	3	90.39007	38.2498	101.9953375	1.925340035	1
10	3	0	4	97.36186	40.98785	115.3387379	1.387182403	1
11	3	0	1	70.50691	34.87651	897.417696	0.207656673	1
12	5	0	1	23.37122	32.75673	7.738360804	0.241186441	0
13	3	0	2	72.40422	34.29383	12.34881707	0.187077158	1
14	3	0	2	81.88275	39.58019	57.40136182	0.624418877	1
15	3	0	3	64.19791	32.33004	33.44798945	0.207656673	1
16	3	0	2	72.52388	36.78358	9.491740224	0.492537268	1
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

위의 과정들을 거쳐서 4가지의 기계학습 알고리듬 - 의사결정 나무, k-최근접 이웃, SVM, 인공신경망-의 학습 모델을 생성하였다.

먼저 의사결정 나무 알고리듬의 경우, 본 연구에서는 의사결정 나무의 알고리듬 중 CART 알고리듬을 사용하였다. CART 알고리듬은 의사결정 나무와 동의어로써 사용될 정도로 가장 널리 쓰이는 알고리듬이며, 후보 나무들을 여러 개 생성하고 그중에서 최적의 나무를 찾아내어 사용한다는 장점이 있다. 또한, 다른 알고리듬들에 비해 연산 속도가 빠르다는 장점도 있다. 본 연구에서는 CART 알고리듬의 매개 변수(hyper parameter)를 complexity parameter(CP)는 0.004, Minsplit은 4, 그리고 Minbucket은 2로 조정하였다. Minsplit은 분할을 시도할 노드에 있어야 하는 최소 관측값의 개수를 의미하며, Minbucket은 터미널 리프 노드의 최소 관측값의 개수를 의미한다. 이 값들은 실험적으로 결정되었다.

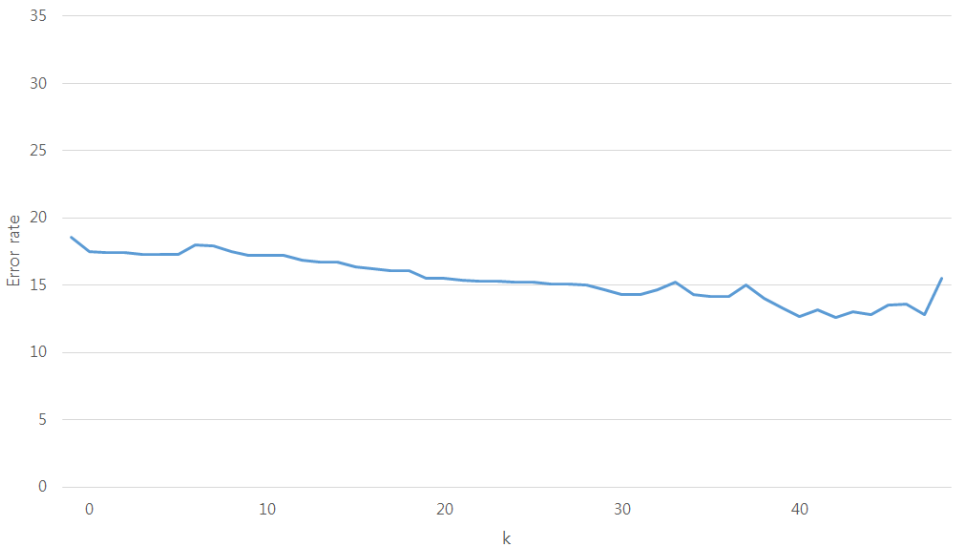
기계학습 알고리듬이 적용되려면 하이퍼 파라미터(hyper parameter)의 조절이 필수적으로 필요하다. 현재까지의 연구들에서는 이러한 하이퍼 파라미터를 최적화하는 과정은 정해진 공식이 없는 상태이다. 적절한 하이퍼 파라미터는 데이터와 모델에 따라 달라지기 때문에 연구자가 실험적으로 최적의 값을 찾아야 할 필요성이 있다. 그러나 우선 모델이 결정된 후부터는 추가적인 세부 변수의 결정 등을 하지 않아도 모델을 쉽게 사용할 수 있게 된다. 결정된 하이퍼 파라미터에 의해 생성된 의사결정 나무는 <그림 3-1>과 같다.



<그림 3-1> 건물에 대해 생성된 의사결정 나무

의사결정 나무 알고리듬이 다른 기계학습 알고리듬과 비교하여 가지는 장점은 나무의 구조를 통해 어떠한 속성이 분류에 영향을 강하게 주는지 쉽게 파악할 수 있다는 것이다. 또한, 이러한 장점 덕분에 의사결정 나무로부터 얻어진 분류 결과를 사람이 해석하고 사용 가능한 형태로 다시 변환할 수도 있다. 건물에 대한 의사결정 나무 구조를 살펴보면 가장 먼저 건물의 넓이로부터 나무의 노드 분할이 이루어지고 있는 것을 볼 수 있다. 즉, 건물의 삭제 또는 유지 여부를 결정하는 가장 중요한 속성이 건물의 넓이라고 말할 수 있는 것이다. 건물의 넓이를 기준으로 일차적인 분류가 일어난 이후에는 가장 가까운 건물과의 거리가 중요한 요소가 되어 다음 단계의 분류가 시작되며, 이어서 건물의 주기, 건물의 둘레 등이 중요한 요소로 판단되었다.

사용된 두 번째 기계학습 알고리듬인 k-최근접 이웃 알고리듬의 경우, 하이퍼 파라미터인 적절한 k 값과 거리 측정 방법을 먼저 지정해야 할 필요가 있었다. 적절한 k 값은 실험 데이터마다 달라질 수 있으므로 실험적으로 결정할 필요가 있다. <그림 3-2>는 실험 데이터에 대해 k 값의 변화에 따른 오차율을 보여준다. k를 1에서부터 50까지 증가시키면서 실험을 수행한 결과, k 값이 45일 때 가장 낮은 오차율을 보였음을 알 수 있었다. 거리 측정 척도로는 실험 데이터의 특성상 건물의 면적과 둘레와 같이 변수 간의 상관관계가 있는 입력변수들이 사용되었기 때문에 거리 측정 시 변수 내 분산, 변수 간 공분산/상관관계를 모두 고려하는 마할라노비스 거리를 사용하였다.



<그림 3-2> k 값에 따른 오차율 변화(건물)

SVM에서는 커널 변수와 정규화 매개 변수를 설정할 수 있는데, 본 연구에서는 Bergstra and Bengio(2012)가 제안한 임의 탐색 방법을 적용하여 변숫값을 설정하였다. 이 방법을 사용하면 최적의 매개 변숫값을 자동으로 찾아주기 때문에 SVM에서는 앞선 두 알고리듬에 비해 매개 변수의 설정 과정이 더 간단해진다고 볼 수 있다.

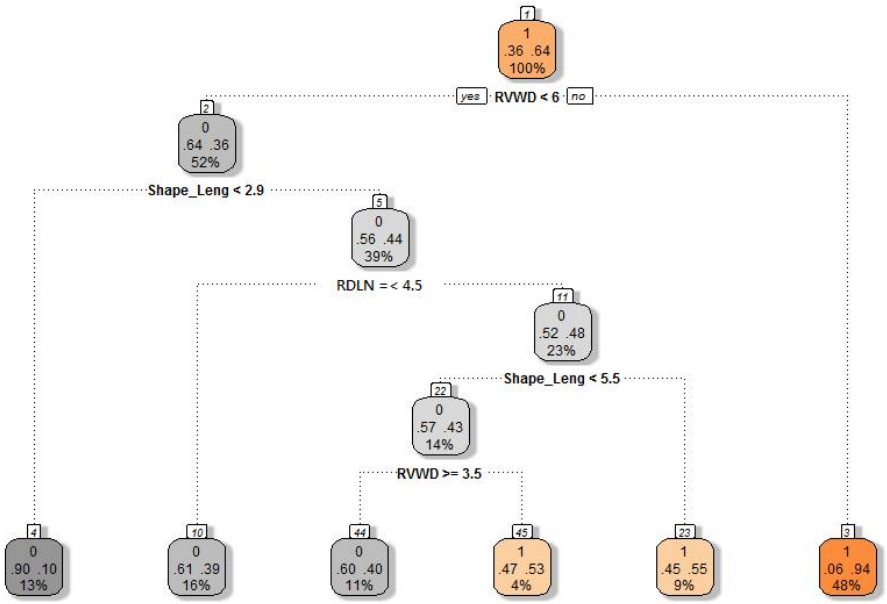
마지막으로 인공신경망은 실험자가 은닉층의 개수, 학습률, 활성 함수의 종류 등을 설정할 수 있다. 본 연구에서는 은닉층의 개수 4개, 학습률 0.1, 활성화 함수 ReLu 함수를 사용하였다. 이 값들은 은닉층의 개수를 2개에서부터 6개까지 조절해 보고 학습률은 0.01, 0.05, 0.1, 0.2, 0.5 등의 값으로 실험해 본 결과 가장 높은 정확도를 보인 값으로 결정하였다. 활성 함수는 Sigmoid 함수와 ReLu 함수를 활용해 실험해 본 결과 Sigmoid 함수 사용 시 예측률이 91.74%로 나타났고 ReLu 함수 사용 시 93.34%로 나타나 ReLu 함수를 선택하였다.

3.1.2. 도로에 적용

도로의 선택적 삭제를 위해서도 건물에서와 마찬가지로 먼저 훈련 데이터 생성이 필요하다. 앞선 전처리 과정과 클래스 부여 과정을 통하여 도로의 선택적 삭제를 위한 훈련 데이터를 생성하였다. 먼저 입력 속성으로는 도로 구분, 포장 재질, 분리대 유무, 차로 수, 도로 폭, 도로 길이, 그리고 중요도를 사용하였다.

출력 속성은 전처리 과정을 거쳐 각 도로객체마다 부여된 0과 1의 클래스를 사용하였다. 앞선 과정들을 거쳐 완성된 입력 속성과 출력 속성을 건물 데이터의 분류 과정에서와 같이 4가지의 기계학습 분류 알고리듬으로 분류하고자 하였다. 본 연구에서는 서로 다른 6명의 지도 제작자가 편집한 지역으로부터 각 20,000개씩의 도로를 추출하여 총 120,000개의 도로 데이터를 사용하여 학습 모델을 생성하였다.

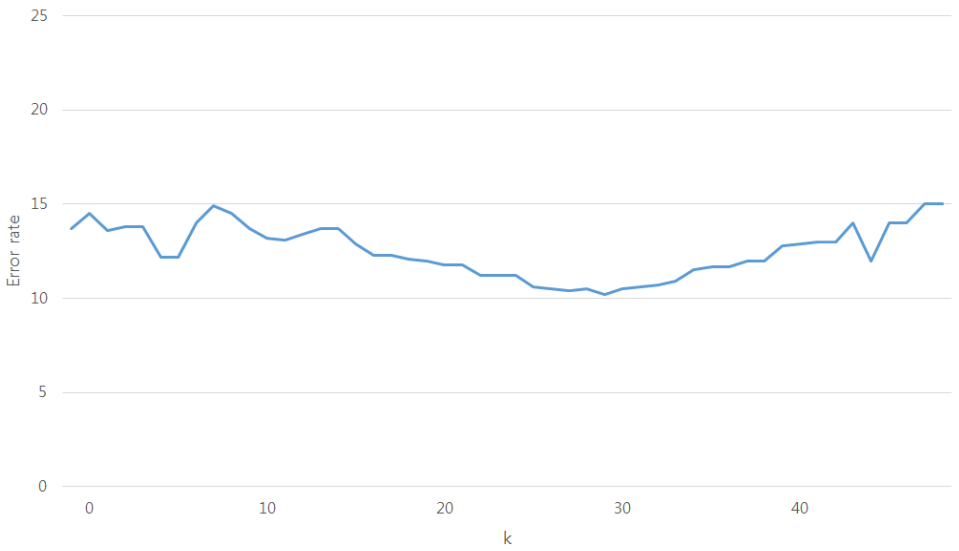
도로의 경우 입력 속성의 개수가 건물의 입력 속성보다 적고, 데이터의 개수 또한 차이가 있으므로 분류 알고리듬들의 세부 변수들의 변경이 필요하였다. 먼저 의사결정 나무 알고리듬의 경우 complexity parameter(CP)는 0.01, Minsplit은 4, 그리고 Minbucket은 2로 결정하였다. CP의 경우 0.01보다 큰 값에서는 학습 속도가 현저하게 저하되고, 0.01보다 작은 값에서는 더는 차이가 발생하지 않았다. Minsplit과 Minbucket은 해당 값보다 더 작아질 경우 학습 속도가 확연하게 저하되고 더 커지면 의사결정 나무 노드의 분할이 너무 낮은 기준으로 진행되어 정확도가 크게 하락하는 문제가 있었다. 도로에 대해 생성된 의사결정 나무는 <그림 3-3>과 같다.



<그림 3-3> 도로에 대해 생성된 의사결정 나무

도로에 대한 의사결정 나무는 가장 먼저 도로 폭 속성으로부터 나무의 노드 분할이 이루어지고 있다. 도로에서는 도로 폭 속성이 도로의 삭제 또는 유지 여부를 결정하는 요소 중에 가장 중요한 속성이라고 해석할 수 있다. 도로 폭에 의해 1차 분류가 이루어진 후에는 도로의 길이, 차로 수의 순으로 순차적으로 분할이 이루어지는 것을 볼 수 있었다. 또한, 도로의 이름이나 도로의 구분 등 의미적인 속성들은 의사결정 나무의 분할에 영향을 주지 못하고 있음을 확인할 수가 있었다.

k-최근접 이웃 알고리듬에 적용에서도 k 값과 거리 측정 방법을 다시 결정해야 할 필요가 있었다. 거리 측정 방법은 변수 간의 상관관계 고려를 위해 마할라노비스 거리를 그대로 사용하였지만, k 값은 1에서부터 50까지의 값을 놓고 별도의 실험을 수행하였다. 그 결과 k 값이 31일 때 가장 낮은 오차율을 보였다(<그림 3-4>).



<그림 3-4> k 값에 따른 오차율 변화(도로)

SVM은 앞서 건물에 적용한 방식이 데이터에 맞는 최적의 매개 변수를 찾아주기 때문에 별도의 조정과정 없이 실험을 진행하였다.

인공신경망은 매개변수들을 조정해 본 결과, 건물에서 사용된 것과 같은 매개 변숫값을 사용했을 때 가장 좋은 결과를 나타냈기 때문에 건물에서 사용된 것과 같은 매개 변숫값 - 은닉층의 개수 4개, 학습률 0.1, 활성화 함수 ReLu - 으로 설정하고 실험을 진행하였다.

3.2. 학습 모델의 생성 및 성능평가

본 연구에서 제안하는 방법의 검증을 위해 먼저 학습 모델을 생성하고 모델의 정확도를 평가하였다. 실험은 Intel i5-4670 3.40Ghz 듀얼 CPU와 RAM 12GB, Windows 10 64bit 환경에서 ESRI 사의 ArcGIS 10.2 소프트웨어를 데이터 처리 과정에서 주로 사용하였으며, R 3.3.0과 Weka 3을 기계학습 적용을 위해 사용하였다. 정확도 평가는 알고리듬별로 정확하게 분류한 건물과 도로객체의 수를 기준으로 진행하였다. 먼저 모델 생성을 위해 사용된 건물 180,000개와 도로 120,000개 데이터로부터 모델의 예측률을 측정하였다. 이 예측률은 건물과 도로의 1:25,000에서의 삭제 여부를 정확하게 예측한 값을 의미하고, 오차 행렬(error matrix)의 형태로 나타내었다. 건물의 분류 결과에 대한 오차 행렬은 <표 3-2>에서 <표 3-5>까지와 같다.

<표 3-2> 의사결정 나무 알고리듬의 오차 행렬(모델-건물)

		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	52,814	6,334
	1-유지됨	7,130	113,722

<표 3-3> k-최근접 이웃 알고리듬의 오차 행렬(모델-건물)

		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	52,369	6,779
	1-유지됨	8,737	112,115

<표 3-4> SVM 알고리듬의 오차 행렬(모델-건물)

		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	53,819	5,329
	1-유지됨	7,531	113,311

<표 3-5> 인공신경망 알고리듬의 오차 행렬(모델-건물)

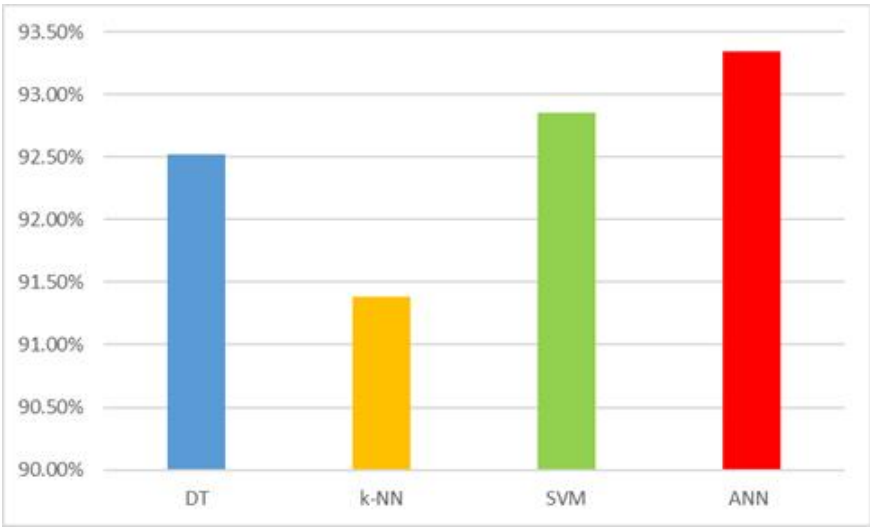
		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	53,070	6,078
	1-유지됨	5,910	114,942

분류 결과를 평가하는 방법으로 가장 대표적인 것은 예측률을 측정하는 것이다. 오차 행렬을 살펴보면 True positive, True negative, False positive, False negative 네 가지로 나눌 수가 있다. 위의 오차 행렬을 예로 들면 실제로 삭제된 건물을 삭제되었다고 예측했거나, 유지된 건물을 유지되었다고 예측한 것이 각각 True positive, True negative이고 삭제된 건물을 유지했다고 예측한 것이 False positive, 유지된 건물을 삭제되었다고 예측한 것이 False negative라고 할 수 있다. 본 연구에서 건물이 남겨졌는지 삭제되었는지가 긍정과 부정의 의미를 담고 있지 않기 때문에 Positive와 Negative는 대치하여 생각할 수도 있다. 이를 활용하여 예측률은 수식 (14)와 같이 구해지게 된다.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad (14)$$

건물에 대한 분류 알고리듬들의 전반적인 예측률은 DT, 92.52%; k-NN, 91.38%; SVM, 92.85%; ANN, 93.34%로 나타났다. 예측률만으로

판단할 경우, ANN이 4가지 분류 알고리듬 중 가장 높은 예측률을 나타냈다(<그림 3-5>).



<그림 3-5> 건물을 대상으로 한 알고리듬별 예측률

예측률만으로는 분류 결과를 제대로 평가하기 어렵다. 데이터의 클래스가 편향되어(bias)있는 경우는 데이터의 양이 많은 클래스가 정확도 전체에 영향을 줄 수 있기 때문이다. 따라서 정밀도(precision)와 재현율(recall), F-measure(F-score)를 각각 구할 필요가 있다. 먼저 정밀도는 수식 (15)와 같이 구한다.

$$\mathrm{Precision} = \frac{TP}{TP+FP} \qquad (15)$$

건물에 대한 분류 알고리듬들의 정밀도는 True positive를 유지된 건물이 유지되었다고 예측한 값이라고 할 때, 각각 DT, 0.9410; k-NN,

0.9277; SVM, 0.9376; ANN, 0.9511로 나타났다.

또한 재현율은 수식 (16)과 같이 계산된다.

$$Recall = \frac{TP}{TP+FN} \qquad (16)$$

건물에 대한 분류 알고리듬들의 재현율은 True positive를 유지된 건물이 유지되었다고 예측한 값이라고 할 때, 각각 DT, 0.9472; k-NN, 0.9430; SVM, 0.9551; ANN, 0.9498로 나타났다. 마지막으로 F-measure(F1-score)가 도출하였다. F-measure는 precision과 recall의 조화평균으로 나타나며, 수식 (17)과 같이 계산할 수 있다.

$$F_1 = \frac{2 \times Recall \times \mathrm{P}recision}{Recall + \mathrm{P}recision} \qquad (17)$$

각 기계학습 분류 알고리듬들의 건물에 대한 F-measure는 각각 DT, 0.9441; k-NN, 0.9353; SVM, 0.9463; ANN, 0.9505로 나타났다.

도로의 대해서도 건물과 마찬가지로 오차 행렬을 작성하고 정확도와 정밀도, 재현율, F-measure를 측정하였다. 먼저 4가지의 기계학습 알고리듬에 의한 도로의 분류 결과에 대한 오차 행렬은 <표 3-6>에서 <표 3-9>와 같다.

<표 3-6> 의사결정 나무 알고리듬의 오차 행렬(모델-도로)

		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	10,501	1,655
	1-유지됨	8,617	99,227

<표 3-7> k-최근접 이웃 알고리듬의 오차 행렬(모델-도로)

		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	10,392	1,764
	1-유지됨	10,008	97,836

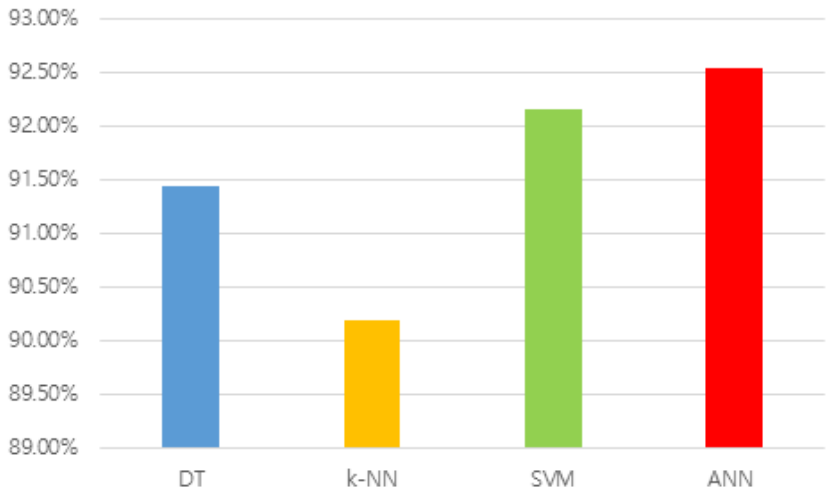
<표 3-8> SVM 알고리듬의 오차 행렬(모델-도로)

		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	10,587	1,569
	1-유지됨	7,851	99,993

<표 3-9> 인공신경망 알고리듬의 오차 행렬(모델-도로)

		예측값	
		0-삭제됨	1-유지됨
실제값	0-삭제됨	10,657	1,499
	1-유지됨	7,453	100,391

도로에 대한 분류 알고리듬들의 전반적인 정확도는 DT, 91.44%; k-NN, 90.19%; SVM, 92.15%; ANN, 92.54%로 나타났다. 전체적으로 건물에서의 값들과 비슷한 양상을 보이었다(<그림 3-6>).



<그림 3-6> 도로를 대상으로 한 알고리듬별 예측률

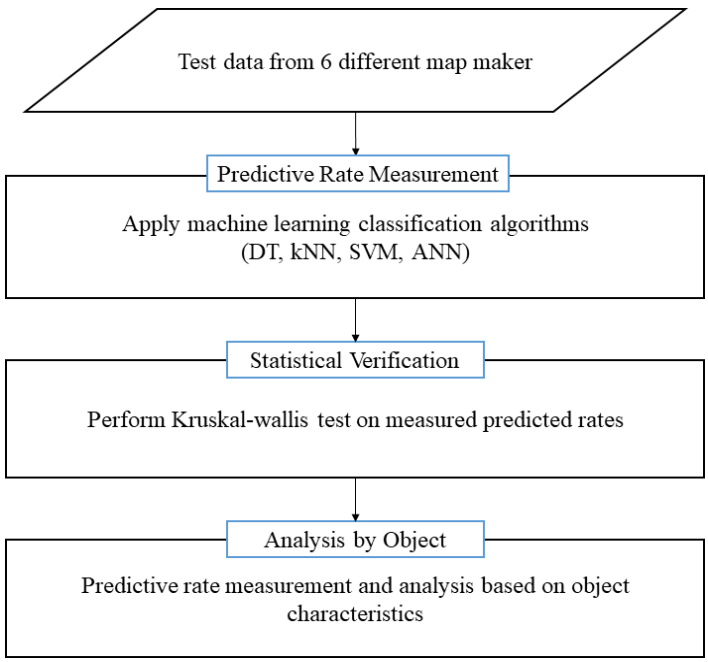
도로도 건물과 마찬가지로 남겨진 도로를 맞게 예측한 것을 True positive라고 정의하고 정밀도와 재현율, F-measure를 계산하였다. 먼저 도로에 대한 분류 알고리듬들의 정밀도는 각각 DT, 0.9201; k-NN, 0.9072; SVM, 0.9272; ANN, 0.9309로 나타났다. 도로에 대한 분류 알고리듬들의 재현율은 각각 DT, 0.9836; k-NN, 0.9823; SVM, 0.9846; ANN, 0.9853으로 나타났다. 각 기계학습 분류 알고리듬들의 도로에 대한 F-measure는 각각 DT, 0.9508; k-NN, 0.9433; SVM, 0.9550; ANN, 0.9573으로 나타났다.

기계학습 기법의 적용을 위해 건물과 도로객체 각각에 대해 4가지 기계학습 알고리듬을 통해 학습 모델을 생성하였다. 생성된 학습 모델들은 건물과 객체 모두 네 가지 알고리즘들이 90% 이상의 예측률을 보였으며 이것은 학습 모델들이 지도 제작자들의 제작 방식을 90% 이상 모사하고 있다고 해석할 수 있다. 학습 모델에서 사용된 학습 데이터는 남겨진 객체(클래스 1)가 삭제된 객체(클래스 0)에 비해 다수의 객체가 존재하는

편향된 데이터라고 할 수 있다. 이 같은 경우에는 예측률만으로는 모델의 성능을 검증했다고 보기 어려우며, 이를 보완하기 위해 정밀도와 재현율, F-measure를 측정하였다. 그 결과 정밀도와 재현율, F-measure 모두에서 0.9 이상의 값을 나타내고 있어 객체 개수의 편향과 관계없이 일정하고 높은 성능을 나타낸다고 해석할 수 있다.

3.3. 모델 적용을 통한 제작자 간 차이의 정량화

앞선 과정을 통해 생성된 기계학습 모델을 서로 다른 6명의 지도 제작자의 축소 편집 결과물에 적용하였다. 모델의 성능평가와 마찬가지로 각 제작자의 축소 편집 결과물 별로 1:5,000에서의 객체가 1:25,000에서 삭제되는지를 예측하여 예측률을 측정하였다. 예측 정확도의 비교를 통해 제작자별로 존재하는 편차를 정량화하는 한편, 편차의 양상을 살펴보기 위해 객체의 특징에 따라 예측률을 따로 측정하여 객체의 나타나는 제작자별 편차의 양상을 살펴보고자 하였다. 이를 위해 건물의 경우 건물의 면적을 5단계로 나누어 단계별 예측률을 측정하였다. 도로의 경우에는 도로 폭을 5단계로 나누어 단계별 예측률을 측정하였다. 모델 적용을 통한 제작자 간 차이의 정량화 과정은 <그림 3-7>과 같다.



<그림 3-7> 제작자 간 차이의 정량화 과정

3.3.1. 건물에 대한 성능평가

먼저 건물에 대해 생성한 학습 모델을 서로 각 지도 제작자의 편집 결과에 적용하여 각각의 예측 정확도를 측정하였다. <표 3-10>은 모델의 정확도와 각 지도 제작자별로 나타난 예측률의 결과이다.

<표 3-10> 건물의 모델 및 각 지도 제작자별 예측률

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Train Model	92.52	91.38	92.85	93.34
Mapmaker A	88.12	87.09	88.02	88.99
Mapmaker B	86.74	85.36	86.13	88.67
Mapmaker C	89.84	88.93	89.96	90.11
Mapmaker D	89.03	88.01	89.91	90.23
Mapmaker E	89.10	87.43	89.34	89.98
Mapmaker F	88.01	87.12	89.99	90.33

적용된 알고리듬의 결과들이 일관성을 보이는지 검증하기 위해 크루스칼 왈리스 검정(Kruskal and Wallis, 1952)을 통해 통계적 검증을 시도하였다. 크루스칼 왈리스 검정은 주로 의학연구에서 많이 쓰이는 통계분석법으로 서로 독립된 세 개 이상의 집단 간의 중앙값 차이를 검정하는 기법이다. 모집단이 정규분포를 따르지 않을 때 사용할 수 있는 기법이기도 하다. 본 연구에서는 크루스칼 왈리스 검정을 통해 관측된 정확도 값이 기계학습 기법에 따라 유의미한 차이가 있는지, 또는 실험 대상 지역에 따라 유의미한 차이가 있는지 검증하고자 하였다. Zhou and Li(2017)의 연구에서도 도로 선택 문제에 있어서 다양한 지도 학습 방법 간의 차이가 유의미한지 검정하기 위해 크루스칼 왈리스 검정을 수행한

바 있다. 크루스칼 왈리스 검정의 귀무가설(null hypothesis, H_0)은 '6개의 서로 다른 축소 편집 결과물 간의 예측률 차이가 없다.'라고 정의할 수 있으며, 대립가설(alternative hypothesis, H_1)는 '6개의 서로 다른 축소 편집 결과물 간의 예측률 차이가 발생한다.'라고 정의할 수 있다.

건물에 대한 서로 다른 제작자들의 측정된 예측률들의 차이가 유의미한지를 검증하기 위한 크루스칼 왈리스 검정 결과는 <표 3-11>과 같다.

<표 3-11> 건물의 모델과 실험 대상 지역의 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
9.6397	5	0.03267	Yes

검증 결과, P-value가 유의수준인 0.05 미만으로 나타나 귀무가설이 기각되면서 서로 다른 축소 편집 결과물 사이의 예측률 차이가 통계적으로 유의한 수준인 것으로 나타났다.

예측률 간의 비교는 지도 제작자 간의 전체적인 편차를 볼 수 있지만, 편차의 양상을 드러내지는 못하고 있다. 제작자 간의 편차의 양상을 살펴보기 위해 대상 지역의 건물의 면적을 5단계로 나누고 단계별로 따로 예측 정확도를 측정하였다. 면적의 단계는 국토교통부 건축정책과에서 발행한 지역별/면적별 건축물 현황 통계자료를 활용하여 설정하였다(국토교통부, 「건축물통계」, 2019). 건물의 면적에 따른 각 지도 제작자별 예측률은 <표 3-12>에서 <표 3-17>까지와 같다.

<표 3-12> 건물 면적별 예측률 - 제작자 A

단위: 정확도(%)

	100㎡ 미만	100㎡-300㎡	300㎡-1000㎡	1000㎡-10000㎡	10000㎡ 초과
DT	82.12	90.22	97.44	100.00	100.00
k-NN	78.88	91.74	96.91	99.88	100.00
SVM	84.79	92.69	98.36	100.00	100.00
ANN	85.54	93.31	98.67	100.00	100.00

<표 3-13> 건물 면적별 예측률 - 제작자 B

단위: 정확도(%)

	100㎡ 미만	100㎡-300㎡	300㎡-1000㎡	1000㎡-10000㎡	10000㎡ 초과
DT	80.89	91.55	98.23	100.00	99.99
k-NN	79.11	91.13	95.53	100.00	99.99
SVM	84.12	93.66	97.26	100.00	99.99
ANN	84.74	93.77	98.17	100.00	99.99

<표 3-14> 건물 면적별 예측률 - 제작자 C

단위: 정확도(%)

	100㎡ 미만	100㎡-300㎡	300㎡-1000㎡	1000㎡-10000㎡	10000㎡ 초과
DT	83.82	90.73	97.89	100.00	100.00
k-NN	81.99	91.88	96.95	100.00	100.00
SVM	85.33	92.98	98.33	100.00	100.00
ANN	85.67	94.33	98.32	100.00	100.00

<표 3-15> 건물 면적별 예측률 - 제작자 D

단위: 정확도(%)

	100㎡ 미만	100㎡-300㎡	300㎡-1000㎡	1000㎡-10000㎡	10000㎡ 초과
DT	82.77	90.25	97.47	100.00	100.00
k-NN	81.15	91.98	97.02	100.00	100.00
SVM	86.12	93.03	98.12	100.00	100.00
ANN	85.89	93.34	98.54	100.00	100.00

<표 3-16> 건물 면적별 예측률 - 제작자 E

단위: 정확도(%)

	100㎡ 미만	100㎡-300㎡	300㎡-1000㎡	1000㎡-10000㎡	10000㎡ 초과
DT	82.69	91.89	98.11	100.00	100.00
k-NN	79.33	92.77	98.09	100.00	100.00
SVM	85.37	93.55	98.98	100.00	100.00
ANN	85.82	94.89	99.02	100.00	100.00

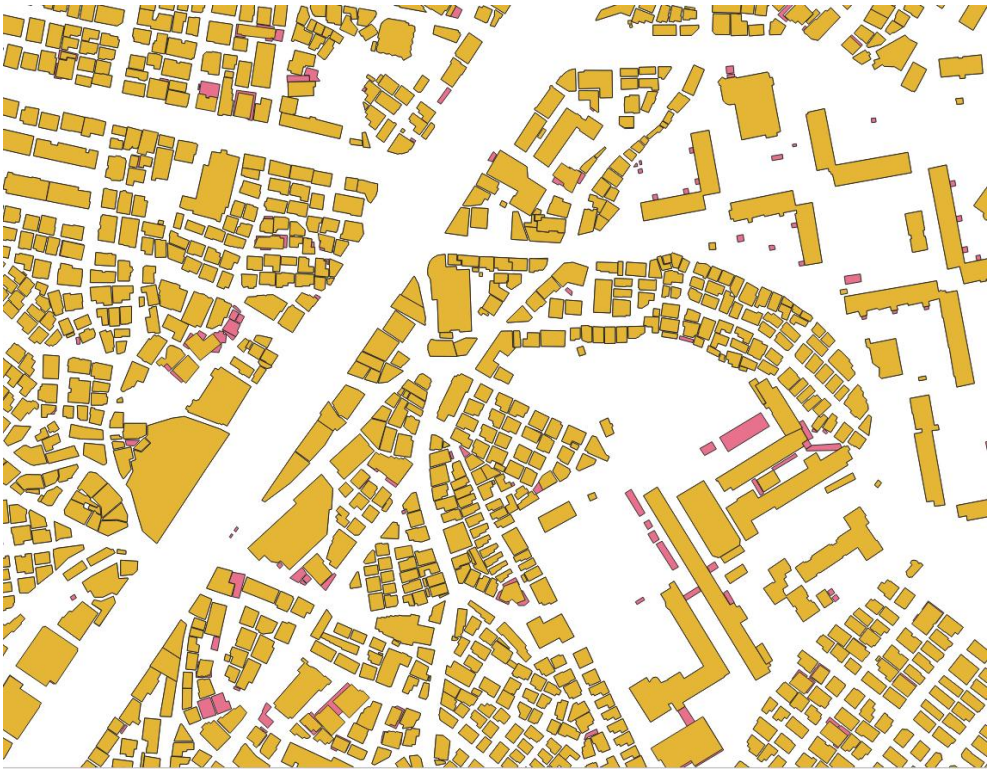
<표 3-17> 건물 면적별 예측률 - 제작자 F

단위: 정확도(%)

	100㎡ 미만	100㎡-300㎡	300㎡-1000㎡	1000㎡-10000㎡	10000㎡ 초과
DT	81.71	90.56	97.38	100.00	100.00
k-NN	80.03	92.05	97.04	100.00	100.00
SVM	85.37	92.99	98.67	100.00	100.00
ANN	85.44	94.11	98.88	100.00	100.00

건물의 면적에 따라 예측률을 측정해 본 결과, 전체 건물의 45%를 차지하는 면적 100m^2 미만의 건물에서 예측 정확도가 낮아지는 동시에 해당 구간에서 제작자들 간의 차이가 주로 발생하고 있음을 볼 수 있었다. 면적이 커질수록 예측 정확도는 상승하고 지도 제작자들 간의 차이는 감소하는 양상을 보였다.

오차의 양상을 더욱 상세히 살펴보기 위해 정성적(시각적) 평가를 수행하였다. 지도 제작자별로 예측된 결과를 시각화하여 비교함으로써, 제작자들 간에 어떠한 차이가 있는지 보기 위함이다.



<그림 3-8> 제작자 A의 건물에 관한 결과 - 음영: 예측 결과의 오답

제작자 A와 제작자 B의 건물 예측 결과의 차이를 시각적으로 살펴보았다. 각 그림에서 색이 다른 건물들은 모델의 예측 결과가 실제 삭제 결과를 예측하지 못한 예측 모델의 오답이다(<그림 3-8>, <그림3-9>). 대체로 소건물에서 오차가 발생하고 있음을 볼 수 있었다. 두드러지는 차이는 대형 건물과 그 대형 건물을 함께 구성하고 있는 부속 건물들에 대해서 차이가 발생했는데, A 제작자의 <그림 3-8>의 중앙 하단에 있는 학교 건물의 부속 건물들에서 모델의 예측과 다른 차이가 발생하였음을 볼 수 있었다. 이는 모델에서는 대형 건물의 부속 건물 같은 경우 필요 때문에 면적이 작은 건물들이라도 남겨두는데, A 제작자의 경우 해당 건물들을 삭제했기 때문에 예측 결과와의 차이가 발생한 것으로 보인

<그림 3-9> 제작자 B의 건물에 관한 결과 - 음영: 예측 결과의 오답

다. B 제작자의 경우 중앙에 있는 대형 건물과 그 부속 건물에서 모델의 예측과 일치한 점으로 보아 의도적으로 해당 건물들을 남겨두었고, 이것이 모델의 예측값과 일치한 결과를 보였다고 생각된다.

특이한 점으로 제작자 B의 경우 10,000m^2 이상의 건물에서 오차가 발생했다는 점이 있다. 이는 모델에서는 10,000m^2 이상의 대형 건물에 대해서는 100% 유지하도록 예측하는 데 비해, 실제로 B 제작자의 편집 결과에서는 10,000m^2 이상의 건물이 삭제된 예도 있기 때문이다(<그림 3-10>). 이는 데이터 자체의 오류 때문에 발생한 오차라고 볼 수 있다.

<그림 3-10> 제작자 B의 10,000m^2 이상 건물에서의 오차 - 음영: 삭제됨

3.3.2. 도로에 대한 성능평가

도로 데이터에 대해서도 건물과 마찬가지로 생성한 학습 모델을 서로 각 지도 제작자의 편집 결과에 적용하여 각각의 예측률을 측정하였다. 표<3-18>는 도로에 대한 모델의 예측률과 각 지도 제작자별로 나타난 예측률의 결과이다.

<표 3-18> 도로의 모델 및 각 지도 제작자별 예측률

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Train Model	91.44	90.19	92.15	92.54
Mapmaker A	87.31	86.42	87.53	89.17
Mapmaker B	86.01	85.12	86.59	87.44
Mapmaker C	88.25	87.77	88.37	89.15
Mapmaker D	87.67	87.03	89.14	89.39
Mapmaker E	87.81	86.56	88.11	89.64
Mapmaker F	86.10	86.97	88.22	88.36

건물에 대한 성능평가에서와 마찬가지로 크루스칼 왈리스 검정을 통해 적용된 알고리듬의 결과들이 일관성을 보이는지에 대한 통계적 검증을 시도하였다. 검증 결과는 <표 3-19>와 같다.

<표 3-19> 도로의 모델과 실험 대상 지역의 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
2.5368	5	0.04687	Yes

검증 결과, P-value가 0.04687로 나타나 유의수준인 0.05 미만이므로 제작자들 간의 차이가 없다는 귀무가설을 기각하게 된다. 따라서 도로에서도 건물과 마찬가지로 서로 다른 축소 편집 결과물 사이의 예측률의 차이가 통계적으로 유의한 수준인 것으로 나타났다.

도로도 건물과 마찬가지로 제작자 간의 편차의 양상을 살펴보기 위해 도로 폭에 따라 5단계로 구분하여 단계별로 예측률을 측정하였다. 도로 폭에 따른 단계 구분은 국토교통부 도로운영과에서 발행한 도로 현황-도로등급별 차로 현황 통계자료를 활용하여 설정하였다. 도로 폭에 따른

각 지도 제작자별 예측률은 <표 3-20>에서 <표 3-25>까지와 같다.

<표 3-20> 도로 폭에 따른 예측률 - 제작자 A

단위: 정확도(%)

	2차로 미만	2-4차로	4-6차로	6-10차로	10차로 초과
DT	86.07	94.89	98.14	100.00	100.00
k-NN	84.22	92.77	98.18	100.00	100.00
SVM	86.26	93.55	98.70	100.00	100.00
ANN	87.88	94.89	99.01	100.00	100.00

<표 3-21> 도로 폭에 따른 예측률 - 제작자 B

단위: 정확도(%)

	2차로 미만	2-4차로	4-6차로	6-10차로	10차로 초과
DT	84.45	94.12	98.12	100.00	100.00
k-NN	83.30	92.59	98.74	100.00	100.00
SVM	85.01	93.20	98.69	100.00	100.00
ANN	85.66	94.74	98.88	100.00	100.00

<표 3-22> 도로 폭에 따른 예측률 - 제작자 C

단위: 정확도(%)

	2차로 미만	2-4차로	4-6차로	6-10차로	10차로 초과
DT	86.99	95.06	98.24	100.00	100.00
k-NN	86.19	94.57	98.38	100.00	100.00
SVM	86.44	94.96	98.77	100.00	100.00
ANN	88.05	94.59	99.14	100.00	100.00

<표 3-23> 도로 폭에 따른 예측률 - 제작자 D

단위: 정확도(%)

	2차로 미만	2-4차로	4-6차로	6-10차로	10차로 초과
DT	86.74	95.03	98.33	100.00	100.00
k-NN	84.69	93.27	98.21	100.00	100.00
SVM	87.13	95.51	98.65	100.00	100.00
ANN	88.99	95.87	99.30	100.00	100.00

<표 3-24> 도로 폭에 따른 예측률 - 제작자 E

단위: 정확도(%)

	2차로 미만	2-4차로	4-6차로	6-10차로	10차로 초과
DT	86.79	94.77	98.07	100.00	100.00
k-NN	84.31	94.37	98.33	100.00	100.00
SVM	86.98	95.12	98.60	100.00	100.00
ANN	88.65	95.53	99.07	100.00	100.00

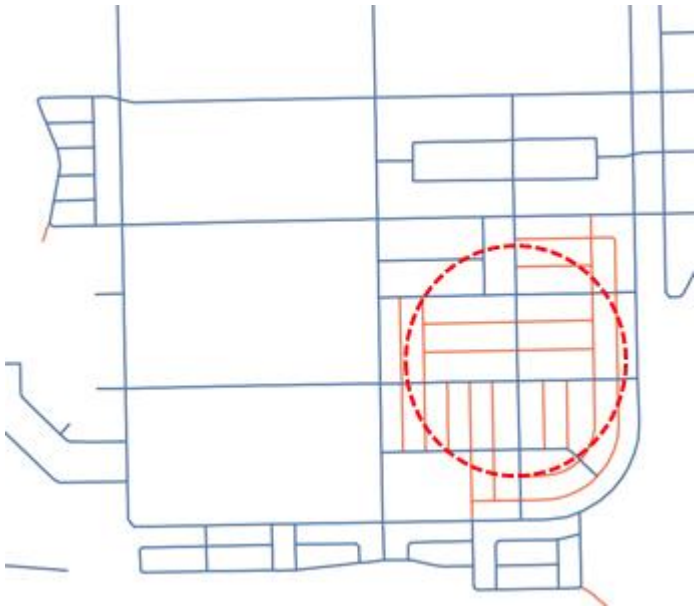
<표 3-25> 도로 폭에 따른 예측률 - 제작자 F

단위: 정확도(%)

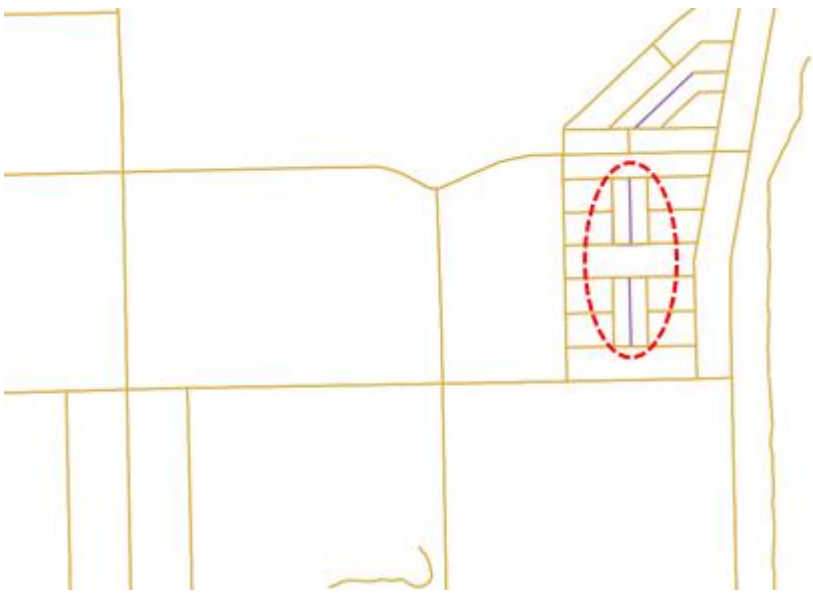
	2차로 미만	2-4차로	4-6차로	6-10차로	10차로 초과
DT	85.04	95.06	98.22	100.00	100.00
k-NN	84.53	93.78	98.31	100.00	100.00
SVM	86.37	94.65	98.66	100.00	100.00
ANN	87.70	95.05	99.01	100.00	100.00

도로 폭에 따라 예측률을 측정해 본 결과, 전체 도로의 68.4%를 차지하는 도로 폭 2차로 미만의 도로에서 예측률이 떨어지는 동시에 해당 구간에서 제작자들 간의 차이가 주로 발생하고 있음을 볼 수 있었다. 도로 폭이 커질수록 예측률이 상승함과 동시에 제작자들 간의 편차가 줄어드는 양상을 보였다.

건물에서와 마찬가지로 오차의 양상을 더욱 상세히 살펴보기 위해 정성적 평가를 수행하였다. 정성적 평가 결과 대표적으로 차이가 나는 부분은 <그림 3-11>, <그림 3-12>와 같다.



<그림 3-11> 제작자 B의 단지 내 도로에 대한 오차 결과



<그림 3-12> 제작자 E의 내 도로에 대한 오차 결과

<그림 3-11>, <그림 3-12>에서 노란색, 보라색으로 표시된 도로는 학습 모델의 예측 결과와 실제 결과가 달랐던 예측 모델의 오차들을 나타낸다. 학습 모델의 예측 결과는 단지 내 도로와 같은 작은 도로의 경우 삭제하는 것으로 판단하고 있다. 그러나 <그림 3-11>의 점선 원으로 표시된 부분에서 볼 수 있듯이 제작자 B의 경우에는 단지 내 도로를 거의 보존하였기 때문에 단지 내 도로에서 학습 모델의 예측에서 벗어난 결과를 보이는 것을 볼 수 있다. 제작자 E의 경우는 일부 단지 내 도로를 보존하기도 하였지만, 학습 모델의 결과와 거의 일치하는 모습을 볼 수 있었다(<그림 3-12>). 이처럼 도로 폭과 길이 모두 작은 소로에서 제작자 간의 격차가 발생하고 있었음을 볼 수 있었다.

3.3.3. 지역 간의 차이에 대한 평가

앞서 측정된 제작자 간의 예측률 차이는 제작자 간의 편차만을 온전히 반영하고 있다고 보기 어렵다. 서로 다른 제작자들이 같은 지역을 대상으로 축소 편집을 한 것이 아니라 제작자마다 다른 지역을 대상으로 축소 편집하기 때문이다. 이러한 차이를 최소화하기 위해 지리적으로 유사한 형태를 가지는 지역을 대상 지역으로 선정하고, 기계학습 모델에 모든 지역의 특성이 반영될 수 있도록 같은 양의 객체들을 임의 추출하였지만, 이것만으로 지역의 차이에 따라 나타나는 변수를 온전히 다 통제했다고 할 수는 없다. 지역 간의 차이가 온전히 통제된 상태로 제작자 간의 편차를 비교하려면 같은 지역에 대해 서로 다른 제작자들이 축소 편집을 한 사례가 있어야 하는데, 이러한 방식으로 축소 편집하기에는 비용과 시간이 비효율적으로 소모되기 때문에 이러한 방식으로 축소 편집이 이루어지고 있는 사례는 없다.

따라서 앞서 드러난 제작자 간 편차에 지역 간의 차이가 얼마나 반영되어 있는지를 측정하여 드러난 제작자 간 편차가 의미 있는 수준인지 점검하는 절차가 필요하다. 이를 위해 같은 제작자가 축소 편집한 서로 다른 지역의 도면을 대상으로 예측률을 측정하여, 그 값 간의 차이가 있는지를 밝혀내고자 하였다. 제작자 간의 편차가 지역의 특성 차이 때문에 발생했다면 같은 제작자가 축소 편집한 서로 다른 지역에서 측정된 예측률들의 차이가 유의미한 것으로 나타날 것이다. 그렇지 않다면, 3.3.2절의 실험 결과는 지역의 특성 차이 때문이 아닌 제작자 간의 숙련도 등으로 인한 편차라고 할 수 있다. 지역 간 차이에 대한 평가는 건물, 도로에서 제작자 간 편차를 측정한 방식과 마찬가지로 훈련된 기계학습 모델로부터 얻어진 예측률 간의 비교를 통해 이루어졌다. 건물과 도로에 대하여 A, B, C 세 명의 제작자가 축소 편집한 세 지역에 대해 각각 예측률을 측정하고, 크루스칼 왈리스 검정을 통해 값 간의 차이가 유의미한 수준인지 검증하였다. 먼저 제작자 A의 건물에 대한 지역 간의 차이는 <표 3-26>과 같다.

<표 3-26> 건물에서의 지역 간의 차이 - 제작자 A

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Area 1	88.52	87.38	87.85	89.34
Area 2	88.33	87.21	88.10	89.55
Area 3	87.98	87.77	87.25	88.98

작업결과 발생한 지역 간의 예측률 차이가 통계적으로 유의미한지 검증하기 위해 앞서 했던 과정과 마찬가지로 크루스칼 왈리스 검정을 수행하였다. 검정 결과는 <표 3-27>과 같다.

<표 3-27> 제작자 A의 건물에 대한 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
0.46154	2	0.7939	No

검정 결과, P-value가 0.7939로 나타나 유의수준인 0.05 이상이므로 지역 간의 차이가 없다는 귀무가설이 채택되게 된다. 따라서 제작자 A가 편집한 서로 다른 지역의 편집 결과에서는 지역 간의 차이로 인한 차이가 발생하지 않았다고 볼 수 있다.

다음으로 제작자 B의 건물에 대한 지역 간의 차이는 <표 3-28>과 같다.

<표 3-28> 건물에서의 지역 간의 차이 - 제작자 B

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Area 1	89.12	87.87	89.75	90.21
Area 2	88.41	87.57	88.77	89.14
Area 3	86.58	85.73	86.25	88.96

제작자 A의 결과와 마찬가지로 B의 결과에서도 크루스칼 왈리스 검정을 수행하였다. 그 결과는 <표 3-29>와 같다.

<표 3-29> 제작자 B의 건물에 대한 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
5.1154	2	0.0775	No

그 결과, P-value가 0.0775로 나타나 유의수준인 0.05 이상이므로 지역

간의 차이가 없다는 귀무가설이 채택되게 된다. 따라서 제작자 B가 편집한 서로 다른 지역의 편집 결과에서 역시 지역 간의 차이로 인한 차이가 발생하지 않았다고 볼 수 있다.

다음으로 제작자 C의 건물에 대한 지역 간의 차이는 <표 3-30>과 같다.

<표 3-30> 건물에서의 지역 간의 차이 - 제작자 C

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Area 1	87.88	86.90	89.03	89.54
Area 2	88.10	87.02	89.52	89.39
Area 3	87.66	87.04	89.77	90.11

제작자 A, B의 결과와 마찬가지로 C의 결과에서도 크루스칼 왈리스 검정을 수행하였다. 그 결과는 <표 3-31>과 같다.

<표 3-31> 제작자 C의 건물에 대한 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
0.5114	2	0.7788	No

그 결과, P-value가 0.7788로 나타나 유의수준인 0.05 이상이므로 지역 간의 차이가 없다는 귀무가설이 채택되게 된다. 따라서 제작자 C가 편집한 서로 다른 지역의 편집 결과에서도 A, B에서와 마찬가지로 지역 간의 차이로 인한 차이가 발생하지 않았다고 볼 수 있다. 이처럼 건물의 경우, 제작자 A, B, C 모두의 결과에서 지역 간의 차이가 통계적으로 유의미하지 않은 수준인 것으로 나타났다. 이것은 건물의 경우 드러난 제

작자 간의 편차에 지역 간의 차이는 의미 있는 수준의 영향을 주지 못하고 있다는 것을 의미한다. 즉, 건물의 경우에는 축소 편집 결과물에서 제작자 간의 차이 때문에 유의미한 차이가 발생하고 있다는 것이다.

다음으로 제작자 A의 도로에 대한 지역 간의 차이는 <표 3-32>와 같다.

<표 3-32> 도로에서의 지역 간의 차이 - 제작자 A

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Area 1	87.11	86.52	87.74	89.06
Area 2	87.01	86.14	86.79	87.33
Area 3	87.84	87.66	88.20	89.14

건물에서와 마찬가지로 도로에서도 지역 간의 예측률 차이가 통계적으로 유의미한지 검증하기 위해 크루스칼 왈리스 검정을 수행하였다. 검정 결과는 <표 3-33>과 같다.

<표 3-33> 제작자 A의 도로에 대한 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
5.5385	2	0.0627	No

검정 결과, P-value가 0.0627로 나타나 유의수준인 0.05 이상이므로 지역 간의 차이가 없다는 귀무가설이 채택되게 된다. 따라서 제작자 A가 편집한 서로 다른 지역의 도로에 관한 결과에서는 지역 간의 차이로 인한 차이가 발생하지 않았다고 볼 수 있다.

다음으로 제작자 B의 도로에 대한 지역 간 차이는 <표 3-34>와 같다.

<표 3-34> 도로에서의 지역 간의 차이 - 제작자 B

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Area 1	89.01	88.53	89.42	90.10
Area 2	88.14	87.88	88.87	89.65
Area 3	86.47	85.09	87.52	88.91

제작자 B의 결과에 대한 크루스칼 왈리스 검정 결과는 <표 3-35>와 같다.

<표 3-35> 제작자 B의 도로에 대한 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
7.2692	2	0.02639	Yes

검증 결과, P-value가 0.02639로 나타나 유의수준인 0.05 미만이므로 지역 간의 차이가 없다는 귀무가설을 기각하게 된다. 따라서 제작자 B가 축소 편집한 지역의 도로에서는 지역 간의 차이 때문에 유의미한 수준의 차이가 발생하고 있다고 볼 수 있다.

끝으로 제작자 C의 도로에 대한 지역 간의 차이는 <표 3-36>과 같다.

<표 3-36> 도로에서의 지역 간의 차이 - 제작자 B

단위: 정확도(%)

	DT	k-NN	SVM	ANN
Area 1	88.30	87.56	89.55	89.67
Area 2	88.12	86.99	89.31	89.37
Area 3	87.52	87.17	89.19	90.07

또한, 제작자 C의 결과에 대한 크루스칼 왈리스 검정 결과는 <표 3-37> 과 같다.

<표 3-37> 제작자 C의 도로에 대한 크루스칼 왈리스 검정 결과*

* H: 검정통계량 df: 자유도

H	df	P-value	Significant
0.7308	2	0.6939	No

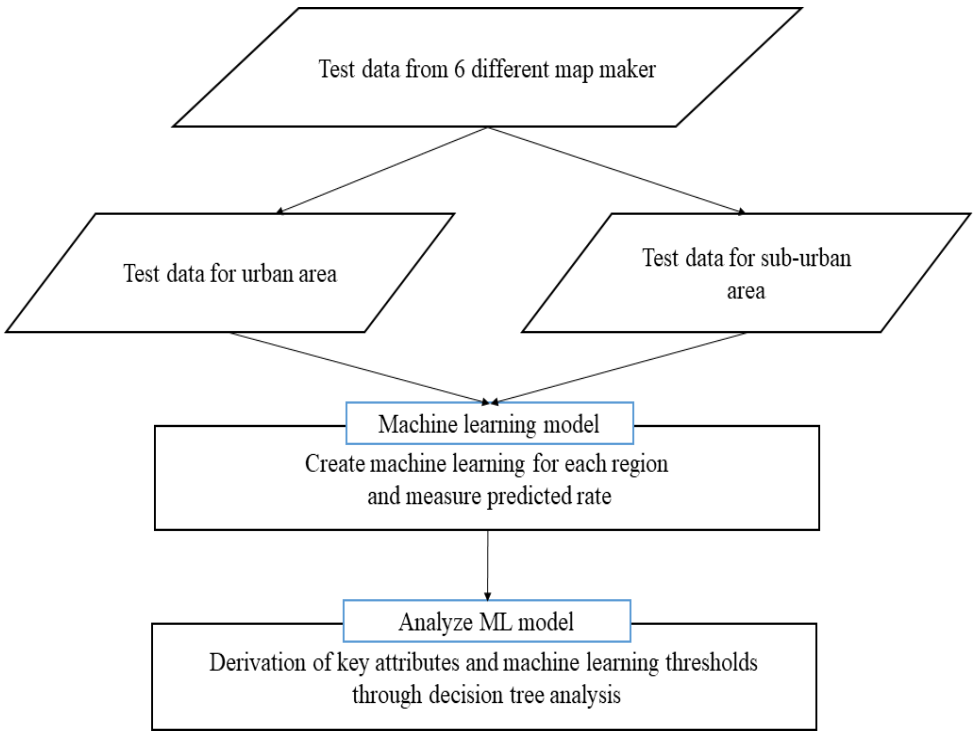
검정 결과, P-value가 0.6939로 나타나 유의수준인 0.05 이상이므로 지역 간의 차이가 없다는 귀무가설이 채택되게 된다. 따라서 제작자 C가 편집한 서로 다른 지역의 도로에 관한 결과에서는 지역 간의 차이로 인한 차이가 발생하지 않았다고 볼 수 있다.

지역 간의 차이가 제작자 간 편차의 결과에 얼마나 영향을 미쳤는지 알아보기 위해 건물과 도로에 대하여 A, B, C 세 명의 제작자가 축소 편집한 세 지역에 대해 각각 예측률을 측정하고, 크루스칼 왈리스 검정을 통해 값 간의 차이가 유의미한 수준인지 검증하였다. 그 결과 건물의 경우, 제작자 A, B, C 모두의 결과에서 지역 간의 차이가 통계적으로 유의미하지 않은 수준인 것으로 나타났다. 그러나 도로의 경우 제작자 B가 편집한 부분에서 지역 간의 차이가 유의미한 수준인 것으로 드러났다. 이는 건물의 선택적 삭제보다 도로의 선택적 삭제가 상대적으로 지역적 특징이 반영될 가능성이 크다고 해석할 수 있다. 또한, 도로객체는 건물객체보다 지역적인 특징에 따라 축소 편집 결과에서 더 많은 차이가 발생하고 있다고 볼 수 있다.

3.4. 도시와 비 도시지역에 적용

앞선 실험들을 통해 현재 제작되고 있는 축소 편집된 수치지형도는 제작자에 따라 유의미한 수준의 차이를 보임을 밝혀냈다. 이는 현재 축소 편집된 수치지형도의 품질이 균질하지 않고, 데이터의 일관성을 담보하지 못하고 있다는 뜻으로 볼 수 있다. 가장 큰 원인은 현존하는 축소 편집 규정의 모호함이라고 할 수 있다. 본 연구에서는 제작자 간의 편차를 밝히는 것에서 나아가 기계학습 기법이 축소 편집 결과물의 품질 향상에 실질적으로 이바지할 방법을 모색해 보고자 하였다.

이를 위해, 훈련 데이터를 도시와 비 도시 지역으로 구분하고 각 지역에 대한 기계학습 모델을 생성하였다. 도시와 비 도시 지역의 구분은 이재빈 등(2018)을 참고하여 “대지” 지목이 전체 면적의 15%를 초과하는 지역을 도시 지역으로 구분하고, 5%미만인 지역을 비 도시 지역으로 구분하였다. 앞선 실험들과 마찬가지로 각 지역에 대해서 기계학습 모델의 예측률을 측정하였다. 나아가 축소 편집 결과물의 품질 향상에 이바지하고 기존 축소 편집 규정의 모호함을 보완할 수 있도록 사용된 기계학습 알고리듬 중 의사결정 나무의 분류 결과를 분석하여 건물과 도로의 선택에 어떤 속성이 얼마나 큰 영향을 주는지 분석하였다. 또한, 의사결정 나무의 노드에서 분화가 일어날 때 사용된 속성값들을 통해 객체의 선택적 삭제 시 기준으로 활용 가능한 수치를 도출할 수 있었다. 본 실험 과정은 <그림 3-13>과 같다.



<그림 3-13> 도시와 비 도시 지역의 기계학습 모델에 대한 분석 순서도

도시와 비 도시 지역에 맞는 기계학습 알고리듬들의 입력변수들을 재설정하였다. 의사결정 나무 알고리듬의 입력변수는 <표 3-38>과 같다.

<표 3-38> 도시와 비 도시 학습모델에서의 의사결정 나무 입력변수

객체/변수	CP (Complexity Parameter)	Minsplit	Minbucket
도시-건물	0.005	4	2
도시-도로	0.008	4	2
비 도시-건물	0.004	4	2
비 도시-도로	0.01	4	2

다음으로 k-최근접 이웃 알고리듬의 입력변수 값들은 <표 3-39>와 같다.

<표 3-39> 도시와 비 도시 학습모델에서의 k-최근접 이웃 입력변수

객체/변수	k	거리 측정 방식
도시-건물	42	마할라노비스
도시-도로	27	마할라노비스
비 도시-건물	45	마할라노비스
비 도시-도로	35	마할라노비스

SVM과 인공신경망의 경우, 기존 학습 모델의 입력변수와 도시와 비 도시 지역에 대해 생성한 학습 모델의 입력변수의 차이가 없었다.

도시와 비 도시 지역에 대해 학습 모델의 예측률을 측정해 본 결과는 <표3-40>과 같다. 도시에서의 예측률이 비도시에서의 예측률보다 건물과 도로 모두에서 조금씩 더 높게 나타나는 것을 볼 수 있었다. 도시이 비도시보다 건물과 도로가 더 규칙적으로 분포하는 양상이 있으므로 이러한 결과가 보이는 것이라고 판단된다.

<표 3-40> 도시와 비 도시에서의 예측률

단위: 정확도(%)

	DT	k-NN	SVM	ANN
도시-건물	93.11	91.47	93.23	93.51
도시-도로	91.87	90.92	92.88	93.12
비도시-건물	91.71	90.80	92.27	92.79
비도시-도로	90.53	90.01	91.87	92.08

생성된 학습 모델 중 의사결정 나무 알고리듬을 통해 생성된 모델은 그 결과의 해석이 가능하다. 의사결정 나무 모델의 결과를 해석함으로써 기계학습 알고리듬을 통해 현존하는 축소 편집 규정의 모호한 부분을 보완하려 하였다. 도시지역의 건물에 대한 의사결정 나무를 분석한 결과, 의사결정 나무의 노드 분할이 '건물의 면적'에서부터 시작되는 것을 볼 수 있었다. 즉 건물의 면적이 선택적 삭제에 있어서 가장 중요한 속성이라고 볼 수 있다. 건물 면적의 표준화된 값을 고려하여 상위 35% 이상이면 선택하고 이하이면 삭제하는 기준이 적용된 것으로 해석할 수 있다. 면적에 의한 분할 이후로는 '가장 가까운 건물과의 거리' 노드 분할이 일어났다. 가장 가까운 건물과의 거리는 20m를 기준으로 노드 분할이 일어났으며, 이를 통해 면적에 의해 삭제되어야 하는 건물 중 20m 이상 떨어져 있는 건물들은 독립된 건물 객체로 판단하여 삭제하지 말아야 한다는 것으로 해석할 수 있었다. 건물의 면적과 건물 간 거리를 기준으로 삭제 대상이 된 건물들은 높이와 둘레 길이에 따라 다시 의사결정 나무의 노드가 분할되는 것으로 나타났다. 높이의 경우 15m 이상, 둘레의 경우 표준화된 값을 고려하여 상위 23% 이상의 건물은 면적 기준으로 삭제되는 건물이라 할지라도 남겨져야 하는 것으로 판단되었다. 이 이외의 속성들은 건물의 선택적 삭제에 큰 영향을 주지 않는 것으로 나타났다.

도시지역의 도로에 대해 생성된 의사결정 나무를 분석한 결과 가장 먼저 도로 폭 속성으로부터 의사결정 나무의 노드 분할이 이루어지고 있다. 도로에서는 도로 폭 속성이 도로의 삭제 또는 유지 여부를 결정하는 요소 중에 가장 중요한 속성이라고 해석할 수 있다. 기준이 되는 도로 폭은 7m로 그 미만의 도로객체에 대해서는 우선으로 삭제가 이루어진다고 볼 수 있다. 도로의 길잇값은 건물의 면적과 마찬가지로 표준화된 값

으로 입력되었기 때문에 상위 27% 이상의 길이를 가진 객체는 남겨진 것으로 나타났다. 도로 폭과 길이에 따라 남겨진 도로 중 차선 수가 2차선 미만인 소로들에 대해서는 삭제된 것으로 나타났다.

비도시 지역의 건물에 대한 의사결정 나무의 결과는 도시지역에서의 결과와 분할되는 속성의 순서는 같았으나, 그 값들에 있어서 차이가 나타났다. 가장 가까운 건물과의 거리 속성에서 도시에서보다 낮은 8m로 노드 분할이 이루어져 있으며, 이는 비도시 지역의 경우 도시지역보다 건물들이 산재하여 있으므로 독립된 건물이라고 볼 수 있는 기준이 상대적으로 덜 엄격해지기 때문이라고 할 수 있다. 건물의 높이 또한 6m 이상의 건물들은 삭제하지 않는 것으로 판단했다. 도시지역보다 전체적으로 낮은 높이의 건물들이 많기 때문이라고 판단된다.

비도시 지역의 도로에 대한 의사결정 나무의 결과 또한 도시지역에서의 결과와 분할되는 속성의 순서는 같고 도로 폭과 차선 수의 경우 기준이 되는 값들도 같았으나 도로 길이 속성의 경우 도시에서보다 큰 값에서 분할이 이루어졌다. 비도시의 경우 도로의 길이는 길지만 도로 폭이 좁거나 차로 수가 작은 도로들이 많아서 도시에서와는 다른 결과가 나온 것으로 판단된다.

분석 결과를 통해 관련 축소 편집 규정「지형도 도식적용규정」(시행 2019. 7. 1)의 개정안 예시를 제안한 내용은 <표 3-41>과 같다.

<표 3-41> 분석 결과에 따른 규정 개정안

현행	개정안
제 89조(도로의 표시 방법) ①~② (생략) ③ 축척별 표현방법은 다음 각 호와 같다. 1. 1:5,000 및 1:10,000 : 건물 등 여타 기호와 접할 경우 그 사이에 0.2 ㎜의 간격을 두고 표시하며 지도상 길이 1㎝ 이하의 것은 생략할 수 있다. 2. 1:25,000 및 1:50,000 가.~다. (생략) 라. 소로 중 마을(부락)과 마을(부락)을 연결하는 도로, 자동차도와 자동차도간을 연결하는 도로, 관광목표, 공장, 또는 광산 등에 도달하는 도로, 산림, 습지 등을 통과하는 주요 도로 등은 전부 표시하고 기타의 것은 독도의 편의를 고려하여 생략할 수 있다. 마. 시가지 지역, 도로 밀도가 높은 지역 등에 위치한 도로는 독도에 필요하다고 인정되는 주요 도로만 표시하고 기타 독도 편의를 고려하여 생략할 수 있다.	제 89조(도로의 표시 방법) ①~② (현행과 같음) ③ (현행과 같음) 1. (현행과 같음) 2. 1:25,000 및 1:50,000 가.~다.(현행과 같음) 라. 소로 중 마을(부락)과 마을(부락)을 연결하는 도로, 자동차도와 자동차도간을 연결하는 도로, 관광목표, 공장, 또는 광산 등에 도달하는 도로, 산림, 습지 등을 통과하는 주요 도로 등은 전부 표시하고 기타의 것은 독도의 편의를 고려하여 생략하되 도시지역의 경우 소로 중 도로 폭 7m 미만의 도로들에 대해서는 삭제한다. 마. 시가지 지역, 도로 밀도가 높은 지역 등에 위치한 도로는 독도에 필요하다고 인정되는 주요 도로만 표시하고 기타의 것은 독도의 편의를 고려하여 생략하되 도시지역의 경우 소로 중 도로 폭이 7m 이상인 도로 중에서 차선수가 2차선 미만인 도로의 경우 삭제한다.
<신설>	바. 도시 지역의 경우 길이 17m 미만의 도로는 삭제하고, 비도시 지역의 경우 길이 32m 미만의 도로는 삭제한다.

제 122조(건물의 취사 선택) ① 건물의 취사 선택은 밀집건물 구역에서는 전체의 형태에 큰 변화가 없는 정도까지 생략할 수 있으나 용도상 필요한 독립건물은 생략할 수 없다. ② 아래 각 호의 건물기호는 독도에 지장이 없는 범위 내에서 우선적으로 표시해야 한다. 1. 특별시청, 광역시청, 도청 2. 시청, 구청, 군청, 읍, 면, 주민센터 3. 경찰서, 지구대 및 파출소, 소방서	제 122조(건물의 취사 선택) ① 건물의 취사 선택은 면적에 따라 삭제할 수 있으며 그 기준이 되는 면적은 도시지역의 경우 12m^2 미만, 비도시 지역의 경우 20m^2 미만으로 한다. ② 아래 각 호의 건물기호는 독도에 지장이 없는 범위 내에서 우선적으로 표시해야 한다. 1. 특별시청, 광역시청, 도청 2. 시청, 구청, 군청, 읍, 면, 주민센터 3. 경찰서, 지구대 및 파출소, 소방서
<신설>	③ 용도상 필요한 독립건물은 생략할 수 없다. 또한, 도시지역의 경우 가장 가까운 건물과의 거리가 20m 이상인 경우에는 특별한 독립건물로 간주하여 삭제하지 않는다. ④ 도시지역의 경우 높이 15m 이상의 건물은 주요 건물로 간주하여 삭제하지 않는다. ⑤ 비도시 지역의 경우에는 ③번 조항에서 기준이 되는 값을 8m로, ④번 조항에서 기분이 되는 값은 6m로 변경하여 적용한다.

3.5. 소결

본 장에서는 건물 180,000개와 도로 120,000개의 데이터로 학습한 학습 모델을 서로 다른 6명의 지도 제작자가 축소 편집한 축소 편집 결과물에 적용하여 각각의 예측률을 측정하였다. 전반적으로 k-최근접 이웃 알고리듬의 예측률이 낮은 편으로 나타났으며 인공신경망 알고리듬이 대체로 높은 예측률을 보였다.

제작자들 간의 예측 정확도 값의 차이가 유의미한 수준인지 판단하기 위해 크루스칼 왈리스 검정을 수행한 결과, 제작자들에 따라 발생한 예측률의 차이는 건물과 도로 모두에서 통계적으로 유의미한 수준인 것으로 나타났다.

객체의 특징에 따른 정확도의 차이를 살펴본 결과 건물의 경우 모든 결과에서 공통으로 면적 100m^2 미만의 건물에서 낮은 예측률을 보이며 또한 제작자 간의 차이도 100m^2 미만의 건물에서 주로 나타나고 있었음을 볼 수 있었다. 면적 1,000m^2 이상의 경우 100%에 가까운 예측률을 보였다. 도로의 경우 주로 삭제 대상이 되는 2차로 미만의 도로가 대부분을 차지하는데(전체 도로의 68.4%) 이 도로들에서 대체로 예측률이 낮아지는 경향을 보이고 또한 제작자 간의 차이가 두드러지게 나타나고 있었다. 전체적인 예측률 분포의 양상은 건물에서의 그것과 유사하게 나타났다.

지역 간의 차이가 제작자 간의 차이에 얼마나 영향을 주고 있는지 알아보기 위해 같은 제작자가 편집한 서로 다른 지역의 축소 편집 결과물에서의 예측률 차이를 분석하였다. 그 결과, 건물에 대해서는 지역의 차이가 예측률에 유의미한 수준의 영향을 주지 못하는 것으로 보였으며, 도로의 경우는 제작자 B의 결과에서 지역의 차이가 유의미한 영향이 있

었지만, 다른 두 제작자의 경우에는 지역의 차이가 유의미한 수준의 영향을 주지 못하고 있음이 드러났다.

본 연구에서 사용된 기계학습 기법의 활용 방안 모색을 위해 지역의 특징에 따라 최적화된 기계학습 모델을 생성하고 해당 모델의 분석을 통해 현존하는 규정의 미비점들을 보완하고자 하였다. 이를 위해 학습 데이터를 도시지역과 비도시 지역으로 구분하여 각각에 대해 기계학습 모델을 생성하여 모델들의 예측률을 측정하는 한편 적용된 기계학습 알고리듬 중에 해석이 가능한 의사결정 나무의 결과를 분석하였다. 그 결과 도시지역의 건물과 도로, 비도시 지역의 건물과 도로의 선택적 삭제에 있어서 어떠한 속성들이 중요한 요소로 작용하는지, 또한 삭제되는지를 결정하는 기준은 얼마인지를 알아낼 수 있었다.

4. 결론 및 고찰

본 연구에서는 기계학습 기법이 지도 일반화 과정에서 사용될 수 있는 활용 방안을 보이고자 하였다. 이를 위해 먼저 기계학습 기법을 활용하여 지도 제작자 간의 차이를 분석함으로써 지도 일반화 과정에서 사람의 개입으로 인해 발생하는 오차를 분석하고자 하였다. 또한, 도시지역과 비도시 지역 각각에 대해서 각 지역에 맞는 기계학습 모델을 생성하고 분석함으로써 객체의 선택적 삭제에 영향을 미치는 속성을 분석하고, 나아가 이를 통해 현재의 축소 편집 규정을 보완하고자 하였다.

이를 위해 1:5,000 수치지형도와 1:25,000 수치지형도를 활용하여 건물 180,000개, 도로 120,000개의 학습용 데이터를 생성하고, 4가지의 기계학습 알고리듬(의사결정 나무, k-최근접 이웃, SVM, 인공신경망)을 학습시켜서 1:5,000 수치지형도에서 1:25,000 수치지형도로의 축소 편집 시에 건물과 도로객체의 선택 여부를 예측할 수 있는 기계학습 모델을 생성하였다. 생성된 학습 모델을 서로 다른 6명의 지도 제작자의 축소 편집 결과물에 적용하여 나타나는 예측 정확도를 측정함으로써 제작자 간에 존재하는 편차를 정량화하고자 하였다. 학습 모델은 건물 180,000개, 도로 120,000개의 데이터를 사용하여 4가지의 기계학습 알고리듬을 학습시켜서 생성하였다. 학습 모델을 생성한 후 6개의 서로 다른 축소 편집 결과물에 적용하여 각각에서 나타나는 예측률을 측정하였다. 그리고 제작자 간에 나타나는 예측 정확도의 편차가 유의미한 수준인지 검증하기 위해 크루스칼 왈리스 검정을 시행하였고, 그 결과 건물과 도로 모두에서 통계적으로 유의미한 편차가 존재함을 볼 수 있었다. 전체적인 정확도 외에도 객체의 특징에 따라 존재할 수 있는 제작자 간의 차이 또한 정량화

하기 위해 건물의 경우 건물 면적에 따라 5단계로, 도로의 경우 도로 폭에 따라 5단계로 나누어 단계별로 나타나는 예측률을 측정하였다.

생성된 모델을 각각의 축소 편집 결과물에 적용하여 예측률을 측정한 결과 건물의 예측률은 최저 85.73%에서 최고 90.88%, 도로의 예측률은 최저 85.44%에서 최고 90.35%로 나타났다. 객체의 특징에 따라서는 건물의 경우 면적 100m^2 미만의 작은 건물들에서 예측 정확률이 떨어지는 동시에 제작자 간의 편차가 있음이 보였고, 건물의 크기가 커질수록 예측률이 높아지는 동시에 제작자 간의 예측률 차이는 거의 보이지 않았다. 도로의 경우 2차로 미만의 좁은 도로에서 제작자의 주관이 많이 개입되고 있음이 보였고, 도로 폭이 넓은 주요 도로일수록 예측률이 높고 제작자 간의 차이는 보이지 않았다. 이를 통해 건물과 도로 모두에서 상대적으로 중요성이 떨어지는 객체들에 대해 지도 제작자들의 주관이 많이 작용하고 있다고 할 수 있다.

드러난 제작자 간의 편차는 지역의 차이를 일부 반영하고 있으므로 지역의 차이가 드러난 편차에 얼마나 영향을 주고 있는지에 대한 분석이 필요하였다. 분석 결과 건물의 경우에는 지역 간의 차이가 제작자 간의 편차에 영향을 주지 못하고 있는 것이 나타났고, 도로의 경우 제작자 B의 경우에 지역 간의 차이가 유의미한 영향을 주고 있음을 보였으나, 다른 제작자들의 경우 지역에 따른 유의미한 차이가 보이지 않았다.

기계학습 기법의 활용을 통해 지도 일반화 품질의 향상에 이바지할 수 있는 활용 방안을 찾고자 하였다. 이를 위해 실험 대상 지역을 도시와 비 도시으로 구분하여 각 지역의 특징에 따라 최적화된 기계학습 모델을 생성하고 해당 모델의 분석을 통해 현존하는 규정의 미비점들을 보완하고자 하였다. 이를 위해 생성된 4가지 기계학습 모델 중 해석이 가능한 의사결정 나무 알고리듬의 결과에 대해 학습 모델을 분석하였다.

그 결과 도시와 비 도시, 건물과 도로객체에 대해서 어떠한 속성이 선택적 삭제에 주로 영향을 미치는지 파악할 수 있었으며, 나아가 선택적 삭제가 어떤 값을 기준으로 이루어지고 있는지도 파악할 수 있었다. 이는 현재 축소 편집 규정의 모호한 점을 보완하는데 활용될 수 있을 것으로 기대된다.

본 연구의 또 다른 의의는 지도 일반화에 기계학습 기법을 적용하기 위해 훈련 데이터 생성 및 훈련 데이터의 입력 속성/출력 클래스의 정의 등 데이터 처리 과정을 보이고 이를 통해 기계학습 알고리듬을 적용하는 과정을 제안했다는 점이다. 벡터 데이터인 지도 데이터를 활용한 연구는 최근의 기계학습 기법을 활용한 연구의 주류를 이루는 이미지 픽셀 기반의 연구와는 차이를 보인다고 볼 수 있으며, 본 연구를 통해 지도 데이터에도 기계학습 기법을 적용하고 활용할 수 있음을 보였다고 할 수 있다.

본 연구의 실용화 및 활용 방안으로는 지도 제작자 간의 편차를 정량적으로 밝혀내는 부분을 고도화함으로써 현재의 지도 축소 편집 결과물의 검수 프로그램으로써 활용 가능할 것으로 생각한다. 또한, 의사결정 나무와 같은 해석 가능한 알고리듬을 더욱 깊게 분석한다면 지도 제작자가 활용 가능한 지도 일반화 일반 규칙을 도출할 수 있을 것으로 기대된다. 학습에 활용된 데이터의 개수를 더 늘리고, 고도화된 다른 알고리듬이나 딥러닝 기법들을 적용한다면 모델의 예측률은 더 상승할 수 있을 것으로 예상한다. 나아가 예측률의 상승은 궁극적으로 현재 막대한 노동력이 요구되는 지도 축소 편집의 과정을 기계가 대체하는 완전 자동화 또한 가능하게 할 것으로 기대해볼 수 있다.

향후 연구로는 본 연구에서 대상으로 한 건물과 도로객체 외에도 POI 등 다른 지도 객체의 선택적 삭제에 대한 자동화가 필요하리라 생각된다. 또한, 의사결정 나무 알고리듬의 고도화를 통해 제안한 규정 개정안

을 보다 구체적으로 보완할 수 있을 것으로 생각한다. 딥러닝 기법들을 적용하여 예측률을 100%에 가깝게 상승시키는 것 또한 지도 일반화의 완전 자동화를 위해 필요한 향후 연구라고 볼 수 있다.

참 고 문 헌

건설기술연구원. (2018). 2018년 건설공사 표준품셈.

국토지리정보원. (2003). 지도 축소 편집 자동화 시스템 개발.

국토지리정보원. (2004). 지도 축소 편집 자동화 시스템 개발.

국토지리정보원. (2006). 지도 축소 편집 자동화 시스템 개발.

국토지리정보원. (2012). 국가기본도 선진화 방안 연구.

국토지리정보원. (2014). 최신 기술을 이용한 지도 제작 체계 개선 연구.

김남신. (2006). 규칙기반 모델링에 의한 지도요소 일반화. *한국학술정보.*

김지영, 허용, 유기윤, 김정옥. (2013). 이종의 공간 데이터 셋에서 매칭 객체 판별을 위한 임계값 산출. *한국측량학회지, 31(1),* 23-28

다다 사토시. (2017). 처음 배우는 인공지능. *한빛미디어*

박경식. (2011). 다축척 수치지도의 도로 및 건물정보 일괄갱신 연구. *한국측량학회지, 29(1)*, 21-28.

박슬아, 유기윤, 박우진. (2014). 수치지도 건물데이터의 매칭 기반 갱신 및 이력 데이터 생성, *한국측량학회지, 32(4-1),* 311-318.

박우진, 이영민, 유기윤. (2013). 수치지형도 일반화를 위한 도로 네트워크 데이터의 선택 기법 연구. *한국측량학회지, 31(3),* 229-238.

박환철. (2000). 수치지도에서 도로 중심선 생성과 보정 기법. *부산대학교 대학원 석사학위논문*

이민부, 김남신, 한균형. (2001). GIS Database 구축을 위한 지형요소의 자동화. *대한지리학회지, 36(2),* 81-92.

이민파. (2001). Map Generalization과 MapGene. 2001년 *한국지도학회 학술발표회,* 74-81.

최병길. (2001). 수치지도 일반화에 있어서 단순화에 관한 연구. *한국측량학회지, 19(2),* 199-208.

최신영, 이성희, 이기준. (1998). 지도 일반화를 위한 위상적 일관성 유지. *한국정보과학회 학술발표논문집.* 158-160.

한균형. (1996). 지도학원론. *민음사.*

Alshehhi, R., Marpu, P. R., Woon, W. L., & Dalla Mura, M. (2017). Simultaneous extraction of roads and buildings in remote

sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing, 130*, 139-149.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46(3),* 175-185.

Anders, K. H., & Sester, M. (2000). Parameter-free cluster detection in spatial databases and its application to typification. *International Archives of Photogrammetry and Remote Sensing, 33(B4/1; PART 4)*, 75-83.

Balboa, J. L. G., & López, F. J. A. (2008). Generalization-oriented road line classification by means of an artificial neural network. *Geoinformatica, 12(3)*, 289-312.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods, 43(1)*, 3-31.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research, 13(1)*, 281-305.

Biljecki, F., Heuvelink, G. B., Ledoux, H., & Stoter, J. (2018). The effect of acquisition error and level of detail on the accuracy of spatial analyses. *Cartography and Geographic Information Science, 45(2)*, 156-176.

Biswajeet, P., & Saro, L. (2007). Utilization of optical remote sensing data and GIS tools for regional landslide hazard analysis using an artificial neural network model. *Earth Science Frontiers, 14(6)*, 143-151.

Buttenfield, B. P. (1991). A rule for describing line feature geometry. *Map generalization: Making rules for knowledge representation,* 150-171.

Chen, J., Hu, Y., Li, Z., Zhao, R., & Meng, L. (2009). Selective omission of road features based on mesh density for automatic map generalization. *International Journal of Geographical Information Science, 23(8)*, 1013-1032.

Choi, J., Oh, H. J., Lee, H. J., Lee, C., & Lee, S. (2012). Combining landslide susceptibility maps obtained from frequency ratio, logistic regression, and artificial neural network models using ASTER images and GIS. *Engineering Geology, 124,* 12-23.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine*

learning, 20(3), 273-297.

Cromley, R. G. (1991). Hierarchical methods of line simplification. *Cartography and Geographic Information Systems, 18(2),* 125-131.

Damen, J., van Kreveld, M., & Spaan, B. (2008, June). High quality building generalization by extending the morphological operators. *In 11th ICA Workshop on Generalization and Multiple Representation, Montpellier, France* (pp. 1-12).

Deng, M., Tang, J., Liu, Q., & Wu, F. (2018). Recognizing building groups for generalization: A comparative study. *Cartography and Geographic Information Science, 45(3),* 187-204.

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization, 10(2)*, 112-122.

Downs, T. C., & Mackaness, W. A. (2002). An integrated approach to the generalization of geological maps. *The Cartographic Journal, 39(2),* 137-152.

Duchêne, C., Baella, B., Brewer, C. A., Burghardt, D., Buttenfield, B.

P., Gaffuri, J., ... & Pla, M. (2014). Generalisation in practice within national mapping agencies. *In Abstracting geographic information in a data rich world* (pp. 329-391). Springer, Cham.

Fischer, M. M., & Leung, Y. (2001). GeoComputational modelling—techniques and applications: prologue. In *GeoComputational Modelling* (pp. 1-12). Springer, Berlin, Heidelberg.

Galanda, M., & Weibel, R. (2002). An agent-based framework for polygonal subdivision generalisation. In *Advances in Spatial Data Handling* (pp. 121-135). Springer, Berlin, Heidelberg.

Gaffuri, J. (2006, June). How to merge optimization and agent-based techniques in a single generalization model. In *workshop on generalisation and multiple representation, Vancouver, United-States.*

Hashemi, M., & Karimi, H. A. (2016). A weight-based map-matching algorithm for vehicle navigation in complex urban networks. *Journal of Intelligent Transportation Systems, 20(6)*, 573-590.

Jiang, B., Liu, X., & Jia, T. (2013). Scaling of geographic space as a universal rule for map generalization. *Annals of the Association of American Geographers, 103(4),* 844-855.

Karsznia, I., & Weibel, R. (2018). Improving settlement selection for small-scale maps using data enrichment and machine learning. *Cartography and Geographic Information Science, 45(2),* 111-127.

Kia, M. B., Pirasteh, S., Pradhan, B., Mahmud, A. R., Sulaiman, W. N. A., & Moradi, A. (2012). An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia. *Environmental Earth Sciences, 67(1)*, 251-264.

Kilpeläinen, T. (2000). Knowledge acquisition for generalization rules. *Cartography and Geographic information science, 27(1),* 41-50.

Lagrange, F., Landras, B., & Mustiere, S. (2000). Machine learning techniques for determining parameters of cartographic generalisation algorithms. *International Archives of Photogrammetry and Remote Sensing, 33,* 718-725.

Lamy, F., Bolte, J., Santelmann, M. & Smith, C. (2002). Development and evaluation of multiple objective decision making methods for watershed management planning. *JAWRA Journal of the American Water Resources Association, 38(2),* 517-529.

Lang, T. (1969). Rules for the robot draughtsmen. *The Geographical Magazine, 42(1),* 50-51.

LeCun, Y., Cortes, C., & Burges, C. J. (2010). Mnist handwritten digit database. *AT&T Labs.*

Leitner, M., & Buttenfield, B. P. (1995). Acquisition of procedural cartographic knowledge by reverse engineering. *Cartography and Geographic Information Systems, 22(3)*, 232-241.

Li, X., & Yeh, A. G. O. (2002). Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science, 16(4),* 323-343.

Li, Z., & Su, B. (1995). From phenomena to essence: envisioning the nature of digital map generalisation. *The Cartographic Journal, 32(1)*, 45-47.

Li, Z. (2007). Digital map generalization at the age of enlightenment: a review of the first forty years. *The Cartographic Journal, 44(1),* 80-93.

Li, Z., & Choi, Y. H. (2002). Topographic map generalization: association of road elimination with thematic attributes. *The Cartographic Journal, 39(2),* 153-166.

Li, Z., Yan, H., Ai, T., & Chen, J. (2004). Automated building

generalization based on urban morphology and Gestalt theory. *International Journal of Geographical Information Science, 18(5),* 513-534.

Mackaness, W. A., Ruas, A., & Sarjakoski, L. T. (Eds.). (2011). *Generalisation of geographic information: cartographic modelling and applications.* Elsevier.

McMaster, R. B.(1991). Conceptual frameworks for geographical knowledge. In *Map Generalization: Making Decisions for Knowledge Representation* (pp. 21-39). Longman's Inc.

Modiri, M., Mohebbi, M., Masoumi, M., Khanlu, H., & Eftekhari, A. (2014). Planimetric features generalization for the production of small-scale map by using base maps and the existing algorithms. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 40(2),* 197-201.

Müller, J. C., Lagrange, J. P. and Weibel, R. (Eds.). (1995). *GIS and generalization: Methodology and practice.* Taylor and Francis: London

Naik, N., & Purohit, S. (2017). Comparative study of binary classification methods to analyze a massive dataset on virtual

machine. *Procedia computer science, 112,* 1863-1870.

Neun, M., Weibel, R., & Burghardt, D. (2004, August). Data enrichment for adaptive generalisation. In *ICA workshop on Generalisation and Multiple representation* (pp. 20-21).

Nogueira, K., Penatti, O. A., & Dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition, 61,* 539-556.

Perkal, J. (1966). An attempt at objective generalization. *Michigan Inter-University Community of Mathematical Geographers, Discussion Paper, 10.*

Peto, M., Kloczkowski, A., Honavar, V., & Jernigan, R. L. (2008). Use of machine learning algorithms to classify binary protein sequences as highly-designable or poorly-designable. *BMC bioinformatics, 9(1),* 487.

Pijanowski, B. C., Tayyebi, A., Doucette, J., Pekin, B. K., Braun, D., & Plourde, J. (2014). A big data urban growth simulation at a national scale: configuring the GIS and neural network based land transformation model to run in a high performance computing (HPC) environment. *Environmental Modelling & Software, 51,* 250-268.

Pilehforooshha, P., & Karimi, M. (2019). An integrated framework for linear pattern extraction in the building group generalization process. *Geocarto International, 34(9),* 1000-1021.

Pradhan, B., Lee, S., & Buchroithner, M. F. (2010). A GIS-based back-propagation neural network model and its cross-application and validation for landslide susceptibility analyses. *Computers, Environment and Urban Systems, 34(3),* 216-235.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning.* Elsevier.

Rangayyan, R. M., Guliato, D., de Carvalho, J. D., & Santiago, S. A. (2008). Polygonal approximation of contours based on the turning angle function. *Journal of Electronic Imaging, 17(2),* 023016. https://doi.org/10.1117/1.2920413

Regnauld, N. (1996, August). Recognition of building clusters for generalization. In *Proceedings of the 7th International Symposium on Spatial Data Handling* (Vol. 1, pp. 185-198). Delft, the Netherlands.

Reumann, K. and Witkam, A. P. (1974), Optimizing curve segmentation in computer graphics, In *Proceedings of International Computing Symposium* (pp. 467-472). American

Elsevier, New York, USA.

Ruas, A., & Plazanet, C. (1996, August). Strategies for automated generalization. *In Proceedings of 7th International Symposium on Spatial Data Handling* (pp. 1-18). International Office for Cadastre and Land Records.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development, 3(3),* 210-229.

Schaffer, G., Peer, M., & Levin, N. (2016). Quantifying the completeness of and correspondence between two historical maps: a case study from nineteenth-century Palestine. *Cartography and Geographic Information Science, 43(2),* 154-175.

Shoman, W., & Gülgen, F. (2017). Centrality-based hierarchy for street network generalization in multi-resolution maps. *Geocarto International, 32(12),* 1352-1366.

Stoter, J., van Smaalen, J., Bakker, N., & Hardy, P. (2009). Specifying map requirements for automated generalization of topographic data. *The Cartographic Journal, 46(3),* 214-227.

Stoter, J., Post, M., van Altena, V., Nijhuis, R., & Bruns, B. (2014). Fully automated generalization of a 1: 50k map from 1: 10k data. *Cartography and Geographic Information Science, 41(1),* 1-13.

Su, B., Li, Z., Lodwick, G., & Müller, J. C. (1997). Algebraic models for the aggregation of area features based upon morphological operators. *International Journal of Geographical Information Science, 11(3)*, 233-246.

Su, B., Li, Z. L., & Lodwick, G. (1998). Algebraic models for collapse operation in digital map generalization using morphological operators. *Geoinformatica, 2(4)*, 359-382.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*(pp. 53-61). MIT press.

Svensson, P. (2016). *Machine learning techniques for binary classification of microarray data with correlation-based gene selection.* [Master's thesis, Uppsala University]. Diva-portal.

Töpfer, F., & Pillewizer, W. (1966). The principles of selection. *The Cartographic Journal, 3(1),* 10-16.

Wang, P., & Doihara, T. (2004). Automatic generalization of roads

and buildings. *Triangle, 50(2),* 1-7.

Ware, J. M., & Jones, C. B. (1998). Conflict reduction in map generalization using iterative improvement. *GeoInformatica, 2(4),* 383-407.

Wei, Z., Guo, Q., Wang, L., & Yan, F. (2018). On the spatial distribution of buildings for map generalization. *Cartography and Geographic Information Science, 45(6),* 539-555.

Weibel, R. (1995). Map generalization in the context of digital systems. *Cartography and Geographic Information Systems, 22(4)*, 259-263.

Weibel, R., & Jones, C. B. (1998). Computational perspectives on map generalization. *GeoInformatica, 2(4),* 307-314.

Wilmer, J. M., & Brewer, C. A. (2010, November 15-19). *Application of the radical law in generalization of national hydrography data for multiscale mapping.* In Proceedings of the Special Joint Symposium of ISPRS Technical Commission IV & AutoCarto in Conjunction with ASPRS/CaGIS, Orlando, FL, United States.

Wu, X., Zhang, H., Xu, Y., & Yang, J. (2017). A comparative study

of various properties to measure the road hierarchy in road networks. *In Spatial Data Handling in Big Data Era* (pp. 157-166). Springer, Singapore.

Vetter, A., Wigley, M., Käuferle, D., & Gartner, G. (2015, August). The automatic generalisation of building polygons with arcGIS standard tools based on the 1: 50,000 Swiss National Map Series. In *Proceedings of the 18th ICA Workshop on Generalisation and Multiple Representation,* (pp.1-12).

Visvalingam, M., & Whyatt, J. D. (1993). Line generalisation by repeated elimination of points. *The cartographic journal, 30(1)*, 46-51.

Zhang, M. (2009). *Methods and implementations of road-network matching,* Doctoral dissertation, Technische Universität München]. Deutsche National Bibliothek.

Zhao, Z., & Saalfeld, A. (1997). Linear-time sleeve-fitting polyline simplification algorithms. In *Proceedings of AutoCarto 13, (pp. 214-223), ACSM/ASPRS'97 Technical Papers*

Zhang, X., Luo, G., He, G., & Chen, L. (2017). A multi-scale residential areas matching method using relevance vector machine and active learning. *ISPRS International Journal of*

Geo-Information, 6(3), 70.

Zhou, Q., & Li, Z. (2014). Use of artificial neural networks for selective omission in updating road networks. *The Cartographic Journal, 51(1),* 38-51.

Zhou, Q., & Li, Z. (2017). A comparative study of various supervised learning approaches to selective omission in a road network. *The Cartographic Journal, 54(3),* 254-264.

Abstract

A Study on Improvement of Map Generalization Using Machine Learning

-Focusing on Selective Omission of Building and Road Data-

Jaeeun Lee

Department of Civil and Environmental Engineering

The Graduate School

Seoul National University

Currently, 1:25,000 digital maps in Korea are created by editing 1:5,000 digital maps. This editing is a process of making a small-scale map from a large-scale map, and in this process, a map generalization technique is inevitably applied. In the past, the generalization of maps has been mainly based on a geometric generalization method, which is generalized using geometric features of objects, or a rule-based method. Currently, the process of reducing editing in Korea is performed through a kind of rule-based method in accordance with the regulations related to editing. However, there are

many areas where the contents of the regulation book are not specific, so there is much room for the editor's subjective intervention. As the subjectivity of the editor is involved in the editing process, the consistency of the quality of the small scale map cannot be guaranteed, and the quality of generalization depends on the individual competency of the editor.

In recent studies of map generalization, there have been steadily raised problems that such human intervention make the results of the generalization of maps inconsistent. Accordingly, the research flow of map generalization has been progressed toward minimizing human intervention and automating data acquisition and data editing processes to ensure consistency of map generalization quality. However, few studies have been conducted on quantitatively revealed research cases and demonstration cases of how human intervention affects the quality of map generalization. Attempts to suggest ways to utilize are also insufficient. There are also insufficient attempts to suggest ways to supplement existing regulations through analysis of the generalized results of maps.

In this study, the difference in the quality of generalization caused by human intervention is quantified by applying the machine learning method, and furthermore, it is intended to suggest a method for improving the quality of the generalized map by utilizing it. For this, a machine learning model that predicts whether buildings and roads can be selected/deleted when scaled-down from 1:5,000 digital maps to 1:25,000 digital maps is created. Then the predicted rate of

predicting whether to select buildings and road objects for each six different map makers were measured. By analyzing the difference between the measured prediction rate and its pattern, it is revealed that there is a significant difference in the editing method between map makers. Another experiment proposes a method for creating a machine learning model for urban centers and non-urban areas, and setting the appropriate machine learning algorithm settings according to each region's characteristics.

In order to evaluate the performance of the learning model, the prediction rate was measured for buildings and roads for each of the four algorithms, DT, k-NN, SVM, and ANN, used in the learning model. In addition, the predicted rate was measured by applying the generated models to six experimental areas, and the difference between the predicted rates was statistically significant through the Kruskal Wallis test.

In this process, since the difference in accuracy may occur depending on the characteristics of the target region, the accuracy of different regions edited by the same producer was measured and the statistical characteristics were analyzed to determine how much the regional characteristics influenced the differences between producers. As a result, in the case of buildings, the difference in accuracy by region was not statistically significant, but in the case of roads, there was a significant difference in some regions. However, the difference by each producer was greater than the difference by region, and this can be interpreted that the difference by producer was also dominant

for road objects. As a result of qualitative (visual) analysis, it was found that the differences were revealed for each producer in the selection and deletion of lane objects, such as roads, in small buildings in urban areas in buildings.

In addition, in order to find a way to utilize the machine learning technique in the generalization of maps, a machine learning model was generated for urban and non-urban areas respectively, and the prediction rate was measured. As a result through the machine learning technique, it was possible to check the properties that have a major influence in the selection and deletion of objects and the settings required for the object selection, and through this, the machine learning algorithm complements the reduced editing rules in the generalization of the map. And it has been shown that it is possible to make basic use of object selection and deletion through machine learning techniques for each feature.

Machine learning techniques can be applied not only to quantify deviations between map makers, but also to automate the generalization of maps. The method proposed in this study also suggests the possibility of automating selection and deletion in the generalization of maps without human intervention through a learning model that has learned from existing map data.

keywords : digital map, machine learning, map generalization, multi-scale database, updating maps

Student Number **: 2011-30271**

감사의 글

소설 “레미제라블”의 주인공 장발장은 5년형을 선고받고 감옥살이 중 여러 차례 탈옥을 시도했습니다. 결국 그는 탈옥 실패로 19년의 징역을 살고 나오게 됩니다. 저의 대학원 생활도 장발장의 삶처럼 끊임없이 도망치려고 했던 삶이었습니다.

이렇게 도망치려고만 했던 저를 잡아주시고, 이끌어주시고, 끝내 박사 학위까지 받을 수 있도록 지도해 주신 유기윤 교수님, 정말 감사합니다. 덕분에 연구자로서 뿐 아니라 한 사람의 인간으로서 성숙해질 수 있었습니다. 따뜻한 격려와 때로는 날카로운 조언으로 저를 이끌어주신 김용일 교수님께도 깊은 감사를 드립니다. 바쁜 일정에도 불구하고 소중한 시간을 내어 제 논문을 심사해 주시고 논문을 완성할 수 있도록 세심하고 꼼꼼하게 지도해 주신 김의명 교수님, 윤준희 박사님, 박우진 박사님께도 진심으로 감사드립니다. 위원님들과 심사 과정을 통해 학문과 연구를 대하는 마음과 태도를 다시 한번 배울 수 있었습니다.

10년이 넘는 대학원 생활 동안 만났던 수많은 연구실 선후배님들께도 깊은 감사의 마음을 전합니다. 격려가 필요할 때는 따뜻한 격려를, 질책이 필요할 때는 모진 질책을 해 주시고 기쁨과 슬픔을 함께 나눠준 여러분들 덕분에 결국 여기까지 올 수 있었습니다. 힘든 시간 보내는 동안 끊임없이 기도와 응원으로 후원해 주신 맑은샘 광천교회, 고향의 늘사랑 교회 가족분들께도 진심으로 감사드립니다. 학위논문 막바지에 여러모로 배려해 주시고 논문에 집중할 수 있도록 도와주신 건설환경공학부 학부 사무실 선생님들께도 죄송한 마음과 깊은 감사를 전합니다.

제가 이 길을 걸을 수 있도록 이끌어주시고, 저의 오랜 공부를 묵묵히

지켜보시면서 응원해 주신 아버지, 어머니 사랑하고 감사합니다. 덕분에 결점 많은 아들이 무사히 졸업할 수 있었습니다. 분야는 다르지만 같이 대학원 생활 하면서 고생하고 있는 사랑하는 내 동생 재혁이에게 특별히 더 사랑의 마음을 전하고 형의 졸업이 동생의 졸업으로 순조롭게 이어질 수 있길 진심으로 응원합니다. 멀리 광주에서 걱정해주시고 변함없이 격려해 주신 사랑하는 할머니 항상 걱정만 끼쳐드려서 죄송합니다. 이제 걱정 그만 끼쳐드리는 손주 되겠습니다. 할머니, 사랑합니다. 학위한다고 제대로 사위 노릇도 못하는 사위를 오랜 시간 인내해주시고 격려해주신 장인, 장모님께도 정말 감사드립니다. 멀리 네덜란드에서 응원 보내주신 처형과 형님께도 감사의 마음을 전합니다.

힘들어서 모든 소망이 다 끊어질 것 같았던 순간에도 나의 소망이자 빛이 되어준 사랑하는 연서, 희서의 앞날에 이 논문과 학위가 조금이라도 도움이 될 수 있기를 희망합니다. 가장 가까이에서 누구보다 마음고생 하면서도 불편한 내색 한번 안하고 기다려준 사랑하는 아내 예은이에게도 미안함과 깊은 감사, 그리고 사랑의 마음을 전합니다.

하나님께서 계시지 않는다고 존재를 부정하던 그 순간에도 나와 함께 하시고 내 인생 가운데 깊이 역사하시는 하나님께 이 모든 영광 돌려드립니다. 이 모든 것이 제힘으로 된 것이 하나도 없음을 고백합니다.

많은 분들이 그동안 고생 많았다고 말씀해 주셨지만, 윤동주 시인의 고백처럼 저 또한 이 논문이 쉽게 쓰여진 것만 같아서 부끄러울 따름입니다. 여기서 그치지 않고 한 사람의 박사로서 온전히 가족과 사회에 제 몫을 할 수 있도록 더욱 정진하도록 하겠습니다. 감사합니다.

2020년 박사논문을 마무리하며,

이 재 은